GENOMIC VERSATILITY IN THE *BURKHOLDERIA* GENUS: FROM STRAINS TO SPECIES

By

Patrick S. G. Chain

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILSOPHY

Microbiology & Molecular Genetics

2011

ABSTRACT

GENOMIC VERSATILITY IN THE *BURKHOLDERIA* GENUS: FROM STRAINS TO SPECIES

By

Patrick S. G. Chain

DNA sequencing has for years helped illuminate the extent of microbial diversity and their phylogenetic relationship, and more recently breathed life into most scientific disciplines, including microbiology, by allowing a complete understanding of genomic structure, functional potential, and patterns of evolution via whole genome sequencing of many species and strains. While the precise definition of bacterial species is still murky, it is clear that the group of closely related organisms classified as *Burkholderia* encompass a very impressive array of phenotypic properties. Using a multitude of genomic and bioinformatic methods, this thesis explores the similarities and differences at various taxonomic levels among genomes sequenced from this genus, attempts to relate some of these discoveries to other species, and discusses the implications of these findings on broader scientific questions. Pangenomic analyses among species of *Burkholderia* reveal an impressive array of variable genes that are not shared among all member of the genus, while comparative pangenomic analysis among strains of the same species clearly show the conservation of its main chromosome and the accelerated rate of evolution in the accessory chromosomes. A lineage-specific analysis of clinical specimens isolated over the course of 20 years reveals the evolutionary strategy for an organism highly adapted to its niche. It is clear that at all levels, the *Burkholderia* display remarkable genomic diversity, enabled in part by their large multi-replicon genome, and responsible for their amazing phenotypic versatility.

ACKNOWLEDGMENTS

I would like to extend my deep appreciation to a large number of individuals for helping me, encouraging me, and having almost forced me to both enter, and continue my graduate studies in search for this elusive PhD. I would like to begin by thanking my advisor, Professor James (Jim) Tiedje for his interest in my education and career, as my conversations with him while I was at LLNL/JGI were the spark that started me on this journey. I particularly thank him for allowing me to participate in a number of projects, and for his encouragement in pursuing new ideas and techniques. I would also like to thank the members of my committee, who were very understanding of my odd route to this PhD, and were very helpful whenever I stopped by for a chat. I am fortunate to have been introduced to so many colleagues, a fantastic array of faculty, and many outstanding scientific opportunities throughout the last several years! I would also like to thank Jim and Linda Beth for being so warm and welcoming, and have made my stays in East Lansing quite comfortable. The same is true for my friends within the Tiedje lab and within the MMG department, thank you.

There are an innumerable number of friends and colleagues whose critical discussions, support, and encouragement throughout the years have helped propel me forward. I would like to specifically thank my dearest colleagues at LLNL who I call family, for without their help and encouragement, I would never have set out to pursue a PhD in this fashion. I am entirely indebted to Dr. Emilio Garcia, Victoria Lao, and members of my group who allowed me to be successful at both student and professional life. I am similarly grateful for all the support I've received at LANL, as my line management has been more

encouraging than I could have hoped! Finally, I'd like to thank my family for their constant support, and to Krista Reitenga, who made it possible for me to finish, and continues to inspire me daily. Without the combined input and help from all of these friends and colleagues, I'm certain this adventure would have turned out quite differently. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1.
An Introduction to Microbial Genomics.

Parts of this chapter have been published in the following articles:

Chain, P. S. G., V. J. Denef, K. T. Konstantinidis, L. M. Vergez, L. Agullo, V. L. Reyes, L. Hauser, M. Cordova, L. Gomez, M. Gonzalez, M. Land, V. Lao, F. Larimer, J. J. LiPuma, E. Mahenthiralingam, S. A. Malfatti, C. Marx, J. J. Parnell, A. Ramette, P. Richardson, M. Seeger, D. Smith, T. Spilker, W. J. Sul, T. V. Tsoi, L. E. Ulrich, I. Zhulin, and J. M. Tiedje. 2006. Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A 103:15280-7.

Yoder-Himes, D. R., P. S. Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, J. M. Tiedje, and R. Sorek. 2009. Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. Proc Natl Acad Sci U S A 106:3976-81.

Chain, P. S. G., D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. Harrison, S. Highlander, P. Hugenholtz, H. Khouri, C. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. Strausberg, G. Sutton, N. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter. 2009. Genomics. Genome project standards in a new era of sequencing. Science 326:236-7.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk, and J. A. Eisen. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056-60.

Introduction

Microbial genomics was thrust upon the world in 1995 when Venter and colleagues proved that computational assembly of sequencing reads could be performed using a random shotgun approach (16). This first complete catalog of genes from an individual organism ushered in a new era in biology and the development of a myriad of bioinformatic methods to exploit the subsequent stream of genome projects. As improvements and cost reductions in sequencing have allowed the genome determinations of numerous strains of the same species, comparative genomic techniques have also emerged and improved to accommodate detailed questions regarding genome differences, function and evolution.

And yet, despite these fantastic advances, our understanding of the greater microbial diversity remains somewhat limited, though expanded explorations of many different habitats and environments via targeted sequencing of the universally conserved small ribosomal subunit (community profiling) have given us a glimpse of the untapped diversity and unseen majority of microbial life. The newer, so-called Next Generation Sequencing (NGS) instruments are only now beginning to allow us to peer even further into the structural and functional component of microbial diversity (environmental or shotgun metagenomics). NGS is also providing a means to explore gene regulation by sampling messenger RNA and small RNAs (via RNA sequencing or RNAseq) that are expressed under different conditions or by different genotypes, facilitating the reconstruction of metabolic pathways. This in turn enables integration of this data with other genomic, phenotypic, metabolomic, and environmental data, including interacting partners, in order to develop an accurate systems biology model.

The constantly increasing rate of sequencing, however, continues to produce more and more genomic data, and has reached such a feverish pace that computational bottlenecks now routinely arise, including such mundane issues as data storage and transfer. Within this chapter is a focused recapitulation of microbial diversity, an overview of the highly versatile genus of *Burkholderia* and its genotypic and phenotypic diversity, the current state of genomic sequencing and methods used to deal with NGS data, and finally ends with a description of this thesis.

Background

A perspective on microbial diversity

While precise numbers of species of bacteria (and archaea) have been a matter of some debate (e.g. (11, 17, 40, 45)), the conclusion that microbial diversity is vast has not been in question. This is no surprise. Microbial life arose roughly 3.7 billion years ago, has helped create the proper conditions that have allowed plant and animal life to thrive over the past ~600 million years, and has thus spent over 3.0 billion years evolving and inhabiting all available niches without eukaryotic competition (4, 43). With the appearance of eukaryotic life, we have also become another environment within which microbial evolution can continue to occur. Indeed it has recently been estimated that there are ten times more bacterial cells living within or on the human body than there are human cells (1, 3). Given an estimated 500-1000 species that reside within the human microbiome, these encode 2-5 million genes, or roughly 50-100 times the number of genes encoded in the human genome (22).

Our current understanding of microbial diversity has been greatly influenced by our ability to read DNA sequences and perform bioinformatic analysis. Indeed, culture-independent approaches such as community profiling using the small subunit ribosomal RNA (16S rRNA) has truly revolutionized the field of microbial ecology in providing a new and improved method to peer into the composition of a given environment in terms of its community members (23, 38, 42). Thanks to whole genome sequencing, microbiologists have also been able to glean novel insights into the potential of organisms in terms of their metabolic capabilities. In trying to accelerate our understanding of microbial diversity, including discovery of novel protein families and providing "anchors" for metagenomic

studies, some recent efforts such as the Genomic Encyclopedia for Bacteria and Archaea (GEBA) have focused on characterizing many genomes based on their phylogenetic position (51). The genomic sequences from phylogenetically, poorly-represented space is a tremendous resource, yet it is clear that these first 56 genomes are only the tip of the iceberg as measured by 16S rRNA phylogenetic distance (Figure 1.1).

The versatile *Burkholderia*: ultimate multi-host, multi-partner symbionts

In addition to determining the complete breadth of microbial diversity, microbiologists are interested in the depth of diversity within individual lineages. The *Burkholderia* represent a fascinating example, as they are a group of betaproteobacteria of staggering phenotypic diversity, likely enabled by their large and plastic genome. Below are a few examples extracted from discoveries and findings over the past few decades, many of them more recent.

While known primarily for their roles in human, animal and plant disease, the majority of *Burkholderia* species are not known to be pathogenic, but many do live in close association with plants, within the rhizosphere and within the cells of a number of different plant tissue types, and even contribute to plant fitness by enhancing growth or providing resistance to stresses (10, 29-30, 49). Others have redeeming qualities such as the ability to degrade recalcitrant chemical compounds present in crude oil, herbicides, and other man-made pollutants (30, 37, 49). The strain *B. vietnamiensis* G4 for example, expresses a toluene o-monooxygenase that can degrade trichloroethylene, a common groundwater pollutant (35), and *B. xenovorans* strain LB400 has been shown to degrade many polychlorinated biphenol compounds (18).

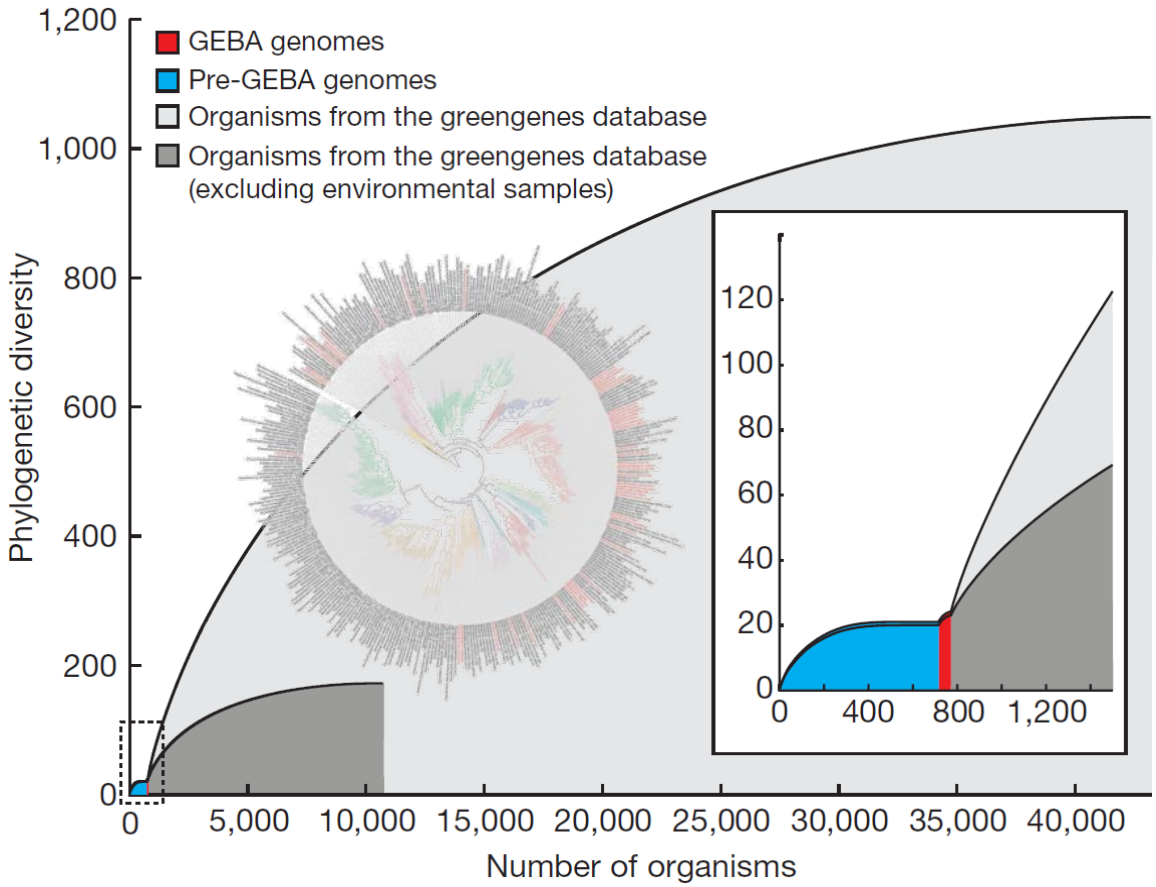


Figure 1.1. Genome sequencing in terms of 16S phylogenetic diversity space. Phylogenetic diversity was measured for the four groups: organisms with sequenced genomes prior to GEBA in blue, the GEBA organisms in red, all cultured organisms in dark grey, and all available non-redundant SSU rRNA genes from the Greengenes 16S database (14) in light grey. The inset box (right) magnifies the first 1,500 organisms. Comparison of the plots shows the phylogenetic 'dark matter' left to be sampled. The inset (circular phylogeny, left) is a maximum-likelihood phylogenetic tree of GEBA genomes and representative bacteria: different phyla are distinguished by color of the branches and GEBA genomes are indicated in red in the outer circle of species names. Modified from Wu et al. (51). For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Burkholderia are the first non alphaproteobacteria shown to nodulate plants. For example, *B. tuberum*, *B. vietnamiensis*, and *B. phymatum* are a few of a growing number of strains that have been shown to induce plant root nodulation in various legume species and are directly involved in the fixation of atmospheric nitrogen, ultimately resulting in increased plant biomass (15, 34, 44, 48). Through the production of many different compounds such as antibiotics, anti-fungals, toxins, iron-chelators, etc., other *Burkholderia* species can benefit plants by outcompeting or suppressing the growth of various bacterial, fungal, protist, and nematode plant pathogens. For instance, the pyrrolnitrin chemical released by a strain of soil-dwelling *B. cepacia* is active against the dry rot fungus *Fusarium sambucinum* that affects potato plants (5). Even the ability of a plant to withstand abiotic stresses can be enhanced by *Burkholderia* strains, as evidenced by the increased growth and physiological activity at low temperatures displayed by grape vines inoculated with *B. phytofirmans* (2). In some cases, the *Burkholderia* endosymbiosis is an obligate interaction for the survival of the plant host. 'Candidatus Burkholderia calva', 'Candidatus Burkholderia nigropunctata' (46), and 'Candidatus Burkholderia kirkii' (47) are endosymbionts of *Psychotria* leaf nodules that are transmitted from one *Psychotria* generation to the next within the plant seed (33). When cured of these leaf-dwelling bacteria in the laboratory, *Psychotria* plants displayed growth deformities and ultimately died (47).

Within this same bacterial genus are many phytopathogenic species, devastating the yield of many important agricultural products. Worse yet, a single species of *Burkholderia* can often infect a wide range of hosts. For example, *B. glumae* causes wilting in rice, tomatoes, sesame, hot pepper, eggplants, perilla, and more than 20 other plants (24), while *B. andropogonis* has the ability to infect mono- and dicotyledonous plant species from 15

different families (reviewed in (9)). Other *Burkholderia* species team with fungi to cause disease. Rhizoxin, a toxin that inhibits plant cell mitosis, is released by a fungal agent of rice seedling blight *Rhizopus microsporus*, but is actually synthesized by its endosymbiont *Burkholderia rhizoxinica* (39). Many other *Burkholderia* also live within phytopathogenic fungi, but do not directly cause disease.

The symbiotic lifestyle of *Burkholderia* also extends to insect hosts. Ants of the species *Atta sexdens rubripilosa* cultivate the fungus *Leucoagaricus gongylophorus* in gardens established by the ants and *Burkholderia* that live in close association with the ants protect the fungal gardens by secreting an antifungal that prevents other fungi from invading the gardens, but is not active against *Leucoagaricus gongylophorus* (41). Other *Burkholderia* species live within the guts of insects, such as those that inhabit *Tetraponera binghami* ants and the cryptic midgutss of broad-headed bugs (family *Alydidae*) (25). Phylogenetic studies performed with the *Burkholderia* community found within the guts of alydid insects placed *Burkholderia* sampled from these insects into a monophyletic clade, suggesting the possibility that coevolution between certain *Burkholderia* strains and the alydid insects has occurred (25).

Finally, certain *Burkholderia* species are notorious for their ability to cause severe infection in humans and other animals. Exposure to *B. pseudomallei*, via contact with contaminated soil or water through cuts, ingestion, or inhalation, may result in melioidosis (12), a disease that presents as either an acute or chronic infection. Meliodidosis is endemic in Australia and some southeast Asian regions and *B. pseudomallei* (13) thus presents a public health threat with the potential to spread by person-to-person and animal-to-person contact. *B. mallei* is closely related to *B. pseudomallei* and causes

glanders in animals including horses, mules, and donkeys and is also able to cause disease in humans through contact with infected animals. Due to historical usage of these *Burkholderia* species to infect animals and people during wartime and their potential for weaponization for future conflicts, both are listed by the US Centers for Disease Control as Category B agents and are Select Agents as declared by the U.S. Department of Health and Human Services and by the U.S. Department of Agriculture because they have the "potential to pose a severe threat to public health and safety".

Individuals with the genetic disorder cystic fibrosis (CF) are at particular risk for contracting severe necrotizing pneumonia infections with species of the *B. cepacia* complex (*Bc*c), and related opportunistic pathogens (such as *B. gladioli*) that prey on the immunocompromised (28). Members of these species persist in the respiratory system and contribute to the decline of CF patient lung function and usually hasten death. Spread of these *Burkholderia* within the CF patient population can quickly lead to devastating epidemics. In addition, the same and other members of the *Bc*c also lurk in industrial settings, where some species contaminate cosmetic and pharmaceutical solutions, disinfectants (19), water supplies, or foods that become poisoned by their toxins (53).

While the use of *Burkholderia* in soil inoculants could hold promise in potential bioremedial agents or as an alternative or supplement to chemical fertilizers or pesticides, caution must be exercised in the development of such applications, given that some of the same species found to enhance plant growth, such as *B. cepacia*, are also well known opportunistic human pathogens. Whether the soil environment serves as a reservoir for human pathogens, and what precise characteristics distinguish harmless environmental

strains from those that cause epidemic infections in clinical environments remain fundamental unanswered questions.

The genomes of various *Burkholderia* were reported to range in size from 4.7 to 9 Mbp and consist of multiple (two to four) circular replicons (8, 50). A large number of insertion sequences have been discovered in many *Burkholderia* genomes (32), which together with the presence of bacteriophage, simple sequence repeats, and variability in copy number of genomic islands have contributed to genomic variations observed in *Burkholderia*, including rearrangements and deletions. Six *Burkholderia* genomes have now been published (6, 20-21, 26-27, 36) and together with the large number of other *Burkholderia* genomes that have been sequenced in the past 7 years, evidence from comparative analyses is mounting that one key to this group's diversity is this highly malleable and large genome. When we compared the *B. xenovorans* LB400 genome to either completed genomes of closely related species, or to other *B. xenovorans* strains via comparative genome hybridization (CGH), it was apparent that many unique LB400 genomic islands were found distributed throughout the genome (Figure 1.3). Interestingly however, genes in chromosome 1 were more conserved in terms of sequence similarity than those found on the other replicons, and upon further investigation, we suggested an accelerated evolutionary rate for these replicons based on relaxed selective pressure for amino acid substitution (6).

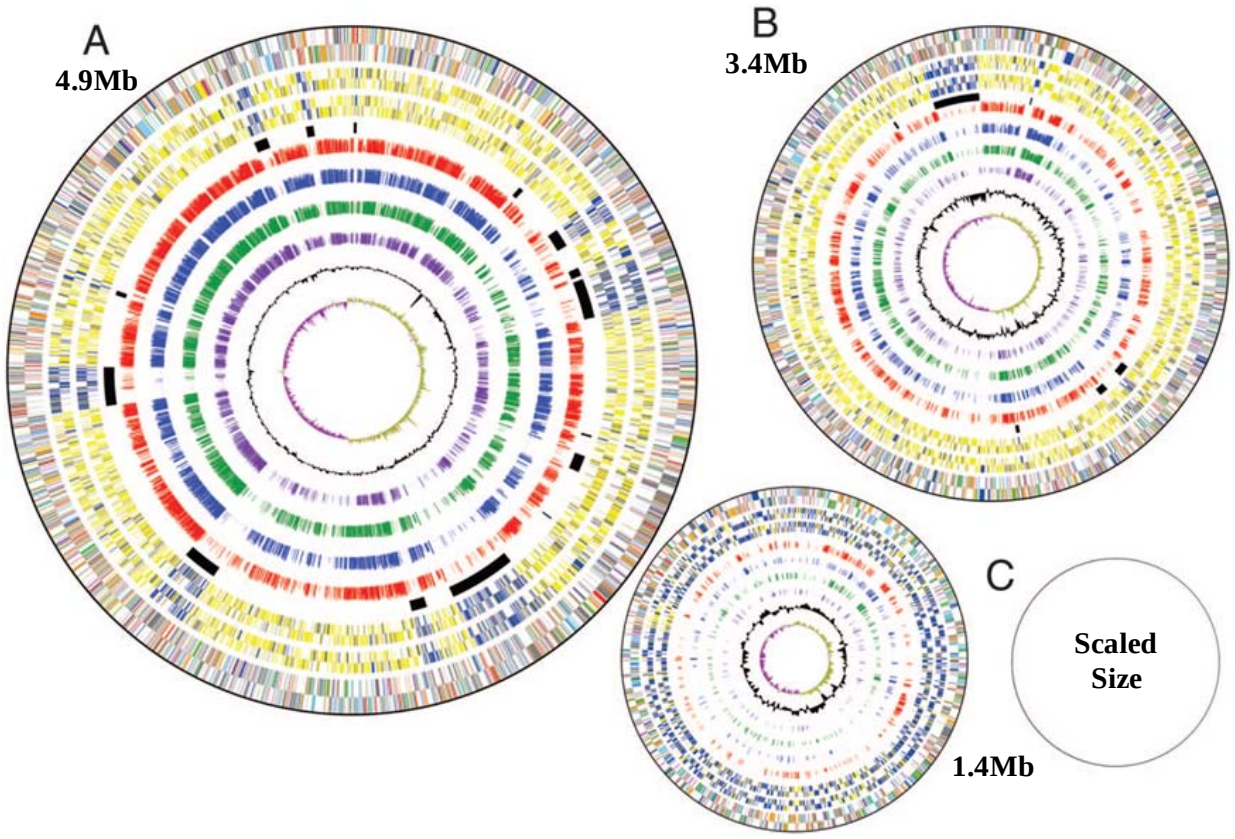


Figure 1.2. Genome of *B. xenovorans* LB400. Circular representation for chromosome 1 (A), chromosome 2 (B), and the megaplasmid (C), scaled to size as indicated. The outer two rings (1 and 2) represent the LB400 genes on the forward and reverse strands, colored by functional class. The next two sets of rings (3-6) represent the CGH data for two other *B. xenovorans* strains (blue = gene absent, yellow = gene present). Ring 7 shows the locations of predicted genomic islands. The next four sets of rings (with bar height relative to the % amino acid identity) are based on reciprocal best BLAST hit analysis (cutoffs: 30% amino acid identity, alignment over at least 70% of the length) between LB400 and *B. cenocepacia* J2315 (red, rings 8 and 9); *B. pseudomallei* (blue, rings 10 and 11); Bcc strain 383 (green, rings 12 and 13); *Ralstonia solanacearum* (magenta, rings 14 and 15). Ring 16 (black) represents GC content and ring 17 represents the G/C skew. Adapted from Chain et al. (6).

Microbial genomics and sequencing technology

Prior to the genome sequencing era, the determination of genetic components responsible for traits of interest was a painstaking and time-consuming process. Microbial genetics was, without knowing it, at the mercy of sequencing technology. Once the potential of this enabling technology was illustrated at the level of whole genome sequencing, many projects of this scale were launched despite the high costs. As sequencing technology improvements took hold, and as more tools, both molecular (to aid in the construction of DNA shotgun libraries) and bioinformatic (for characterization and finishing of genomic sequences) emerged, an exponential number of projects were initiated due to lower costs, creating a sequencing revolution.

While throughput improvements were constantly being made, the Sanger sequencing technology, using fluorescently-labeled dideoxy terminators for laser-based detection of gel-separated DNA amplicons, had been the staple genome sequencing technology for 10 years before the next technology was announced. In 2005, a novel "pyrosequencing" approach from 454 Life Sciences (31) shook the foundation of genomics and it took two years before major sequencing centers warmed to the potential of the many fold higher throughput, yet lower quality, technology. The subsequent widespread adoption of this so-called Next Generation Sequencing (NGS) platform for generating draft sequence, complemented with the higher quality, higher cost Sanger technology opened the door for the Solexa (Illumina) and SOLiD (Applied Biosciences) NGS technologies, released in 2007 and 2008, respectively. These were known as short-read massively parallel platforms capable of producing millions of very short reads (25-36bp at the time of release). These NGS technologies have both contributed to the increased number and pace

of completing microbial genomes, as well as to the growing disparity between drafted and finished genomes ((7), Figure 1.3). These NGS platforms now completely dominate the sequencing market, in part due to a number of novel applications that have been enabled thanks to their cost-effectiveness in generating millions (and now up to billions) of reads (see Table 1.1 below).

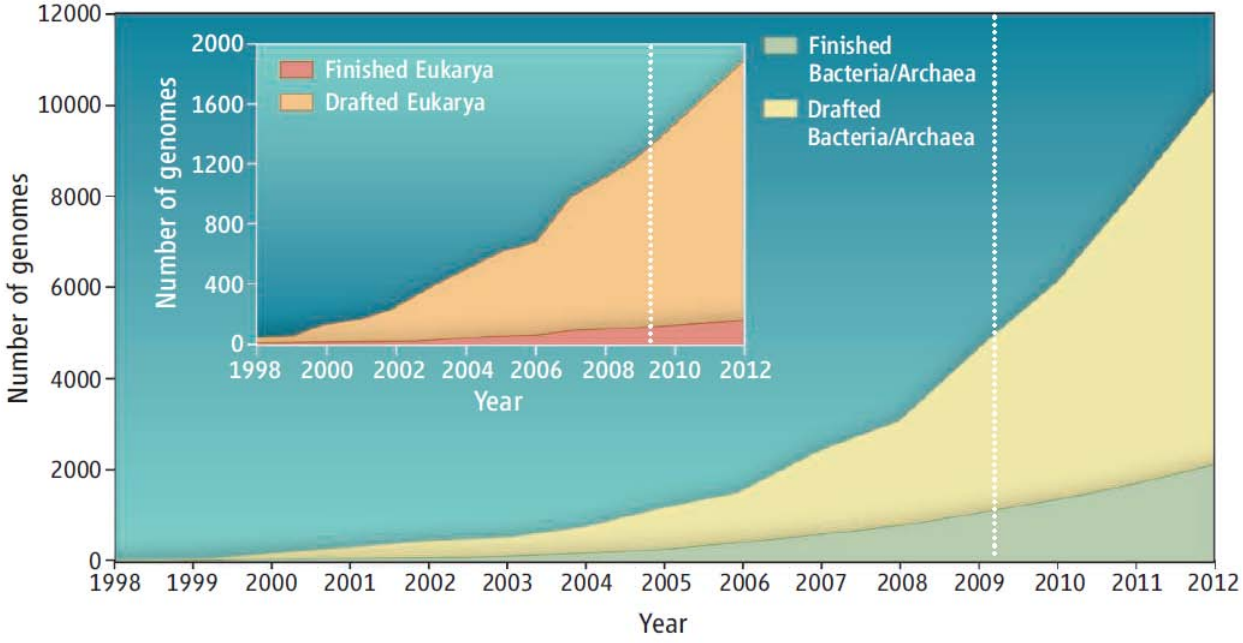


Figure 1.3. Trends in generation of finished and drafted genomes. Figure shows the number of completed and incomplete genomes since 1998, as well as a low end estimated projection for the near-term future. From Chain et al. (7).

While the field of metagenomics was beginning to be explored in a limited fashion, the ability to obtain tens of Mb and now several Gb of sequencing data within days has

completely revolutionized this field (Metagenome report NSF). The throughput of these NGS technologies has also heralded the field of transcriptomic sequencing (also known as digital gene expression or RNA sequencing - RNAseq), where cDNA can be sequenced and the number of reads belonging to a transcriptional unit could serve as a proxy for the level of gene expression. Given a genomic sequence, alignment of sequencing reads to the reference can be used to measure gene expression, as well as to improve gene calling, delineate gene and operon boundaries, and discover small non-coding RNAs.

In the first publication of a bacterial RNAseq-derived expression profile, we leveraged the use of Illumina NGS technology to understand the transcriptional response of two *Burkholderia cenocepacia* strains of the same lineage (52). The two isolates of *B. cenocepacia* (one isolated from soil and another from the sputum of a cystic fibrosis patient) were each grown under conditions mimicking soil and CF sputum, and their transcriptomes sequenced. In this study, while we found differences between the expression profiles derived from cells grown in two different media, we also found significant differences between the global expression profiles of the two strains under the same conditions, despite their high degree of sequence identity, indicating possible strain-specific adaptations to the environmental niche from which they came (Figure 1.4). Illustrating the power of this novel approach, we also discovered at least 13 previously unknown putative non-coding RNAs, 12 of which are preferentially upregulated under soil conditions, suggesting a possible role for ncRNAs when living in the soil environment (52).

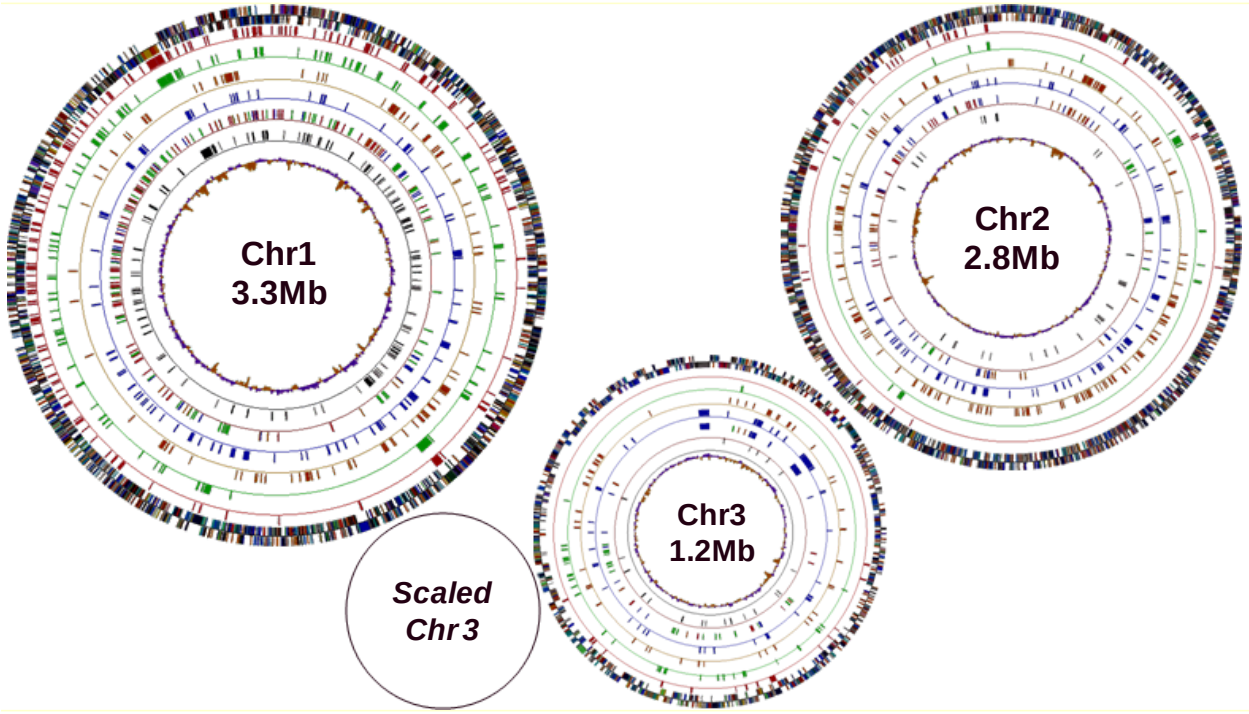


Figure 1.4. Global *B. cenocepacia* differential gene expression comparisons.

Outermost rings indicate all coding regions in the genomes colored according to COG designation. Red and green rings correspond to *B. cenocepacia* strains AU1054 and HI2424 genes induced under cystic fibrosis sputum conditions, respectively. Orange and blue rings correspond to AU1054 and HI2424 genes induced under soil extract conditions, respectively. The brown ring indicates genes differentially regulated between the strains colored by condition of up-regulation. The gray ring indicated genes whose expression is conserved under cystic fibrosis (black) or soil extract conditions (gray) in both strains. The inner most ring shows either high (purple) or low (orange) GC content. Modified from Yoder-Himes et al. (52).

The positive contribution of next generation sequencing technologies to microbial genomics is irrefutable, particularly within the sphere of new applications discussed, such as metagenomics and RNAseq, afforded by short-read sequencing platforms. Unimaginable only a few short years ago, we can now generate several microbial genomes, along with their transcriptomes within a single sequencing run. As the array of competitive and high throughput sequencing platforms continues to expand, we are now witnessing the first single molecule real-time sequencer, the first non-fluorescence based sequencer, the first multiple kilobase sequencer, and possibly in the near future, the first non-enzyme based sequencer (Table 1.1). It took several years to adapt to the current NGS data, including methods to estimate errors and biases within each platform, and to accommodate the shorter read lengths. Investments will have to be made to judiciously integrate these new technologies in current processes, and to explore novel avenues of research enabled by these third generation platforms.

Table 1.1. Available sequencing technologies.

Sequencing Technology	Read Length	Approx. Cost/Mbp	Maximum Bases/Run	Error Rate	Potential Biases/ Comments
Sanger	600-1000	$500	364 Kb	10^{-4} - 10^{-5}	Cloning biases, Random errors
454 GS-FLX or GS-Junior	400	$10 $20	450 Mb 35 Mb	10^{-3} – 10^{-4}	Biases at extreme %GC, Homopolymer errors
Illumina GA$_{II}$X or Hiseq2000	75-125 100	$0.10 $0.05	95 Gb 200 Gb	10^{-2} – 10^{-3}	Biases at extreme %GC
SOLiD 5500 or 5500xl#	75	$0.04	105 Gb 210 Gb	10^{-2} – 10^{-3}	Biases at extreme %GC
Helicos Heliscope	32	$0.45-0.60	37 Gb	10^{-5}	Single molecule, very short reads
Polonator G.007	26	<$0.45	12 Gb	$1:10^{-1}$ – $1:10^{-3}$	Very short reads
Pacific Biosciences PacBio RS#	1-2kb	Not available	60 Mb	$\sim10^{-1}$	Single molecule, real time; strobe sequencing*
Ion Torrent PGM#	~100	~$0.02	10 Mb	10^{-1} – 10^{-2}	Non-light pH detection, Homopolymer errors

*Multiple (2 or more) linked reads, conceptually similar to paired-end reads; #Estimates of output at release of instrument

Bioinformatics: the next revolution

As more and more genomes are sequenced, either finished or in some form of draft, it has become increasingly important to be able to accurately compare them and link differences in genotype to phenotype and to evolutionary history of the organism(s). A

large number of bioinformatics tools and genomic databases have been developed throughout the years to accommodate an increasing number of genomes, a growing number of sequence-based questions, new applications for sequencing such as metagenomics and RNAseq, etc. As we continue to push the utility of sequencing, continue exploring the seemingly unending genomic diversity of microbes, and as new sequencing technologies emerge, a bioinformatics revolution will be required in order to analyze all this data.

Similarity searches lie at the heart of all genomic analyses, be it for general annotation, taxonomical classification of reads or genes, functional assignment of genic or intergenic regions, or for grander whole genome comparative analyses. Many different methods have been developed, but their use is often dependent on the type of data, type of comparison, and type of computational infrastructure utilized. For example, although many early methods were developed specifically for the type of data available when Sanger sequencing was the only technology in use, these do not all work for the types of data generated by NGS platforms. In particular, the sheer volume of data has rendered tasks once deemed trivial for Sanger sequencing, barely feasible. For example, NR database searches using BlastX for either a full run of 454 ($\sim 1 \times 10^6$ reads) or a full lane of Illumina (using the GAiiX, $\sim 3 \times 10^7$ 7reads) would take over 1000 CPU days using a Linux workstation (2.93 GHz Intel Xeon X5570). New methods have yet to be developed to tackle such issues as metagenome assembly using millions to billions of sequence reads, taxonomic or functional classification of millions to billions of reads, short read binning or

annotation, etc. Additional new challenges will emerge when third generation sequencing platforms are fully introduced into the fold.

For completed genomes, the challenge facing researchers is less complex, but still immense: instead of describing a single genome, many genomes are to be analyzed for a large variety of features including an in depth description of their relatedness, or differences that can account for phenotypic variability. Pangenomics is a new field borne out of a growing number of projects whose aim it is to sequence many members of the same species. One of the goals is to define the "core" genome of a species, however many groups utilize different tools for determining the core genes, or use different parameter settings with the same tools. Keeping in mind that methods matter, this thesis explores the genomics of the *Burkholderia* genus, including pangenomics and the use of NGS and associated methods for comparing genomes.

Thesis Outline

In this thesis, I have chosen the *Burkholderia* genus as principal model with which to utilize available, as well as establish novel, tools for comparing genomic sequencing data. There are 25 genomes completed to date, with an additional ~60 other genomes either already drafted or underway. Through the analysis of these genomes, it was clear that many of the incomplete genomes were wrought with sequencing errors and should not be used for detailed comparative analyses. Other completed genomes were found to lack annotations for some of the most conserved genes throughout the entire bacterial phylogenetic tree. For example, RNA polymerase subunit beta (*rpoB*) is labeled a pseudogene in *B. mallei* ATCC 23344 despite having a full length gene and thus is not represented as a protein in the genome. Similarly the translation initiation factor IF-3 is also labeled a pseudogene in *B. cenocepacia* J2315, despite a full length open reading frame. These were erroneously not annotated, and thus extra care was taken to look for mis-annotated regions.

The following chapters are ordered such that the story begins by comparing at a very high level, the gross differences and commonalities between the sequenced species, and ends by extracting the small differences between recently derived isolates of a clonal lineage of one species (the *B. cenocepacia* ET-12 lineage). Thus, Chapter 2 describes an effort to exploit and compare all completed available genomes, which include 25 genomes of 15 species (four species have two to four strains sampled). The aim was to reconcile conflicting phylogenetic inferences, identify genomic features that could define the genus as well as unique features that could define each species or genome and that may be

responsible for organism niche preference, and provide insights into its complex multi-replicon genome.

Chapter 3 examines more closely the relationship between various strains of four species (*B. ambifaria*, *B. cenocepacia*, *B. mallei*, and *B. pseudomallei*), three of which have four sequenced representatives and *B. ambifaria* has two representative strains sequenced. We also take a closer look at the pangenomes of these *Burkholderia* species and try to establish if there are any rules that govern the evolution of *Burkholderia* species, or if there are any constraints on their pangenomes or their replicons. Lastly, Chapter 4 makes use of the speed with which next generation sequencing platforms can generate genomic data, by exploring a number of *B. cenocepacia* ET-12 isolates in a single Illumina sequencing run. This Chapter's focus is even more specific, and is an analysis of NGS technology, bioinformatics methods, and the exploration of four clinically derived ET-12 lineage strains and their comparison to the published J2315 reference sequence. It is clear that at all levels, the *Burkholderia* display remarkable genomic diversity (from strain to species to genus), and this is summarized in Chapter 5.

References

1. 2007. Humans Have Ten Times More Bacteria Than Human Cells: How Do Microbial Communities Affect Human Health?, ScienceDaily, vol. 5 June 2008.

2. **Ait Barka, E., J. Nowak, and C. Clement.** 2006. Enhancement of chilling resistance of inoculated grapevine plantlets with a plant growth-promoting rhizobacterium, Burkholderia phytofirmans strain PsJN. Appl Environ Microbiol **72:**7246-52.

3. **Berg, R. D.** 1996. The indigenous gastrointestinal microflora. Trends Microbiol **4:**430-5.

4. **Brocks, J. J., G. A. Logan, R. Buick, and R. E. Summons.** 1999. Archean molecular fossils and the early rise of eukaryotes. Science **285:**1033-6.

5. **Burkhead, K. D., D. A. Schisler, and P. J. Slininger.** 1994. Pyrrolnitrin Production by Biological Control Agent Pseudomonas cepacia B37w in Culture and in Colonized Wounds of Potatoes. Appl Environ Microbiol **60:**2031-9.

6. **Chain, P. S., V. J. Denef, K. T. Konstantinidis, L. M. Vergez, L. Agullo, V. L. Reyes, L. Hauser, M. Cordova, L. Gomez, M. Gonzalez, M. Land, V. Lao, F. Larimer, J. J. LiPuma, E. Mahenthiralingam, S. A. Malfatti, C. J. Marx, J. J. Parnell, A. Ramette, P. Richardson, M. Seeger, D. Smith, T. Spilker, W. J. Sul, T. V. Tsoi, L. E. Ulrich, I. B. Zhulin, and J. M. Tiedje.** 2006. Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A **103:**15280-7.

7. **Chain, P. S., D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter.** 2009. Genomics. Genome project standards in a new era of sequencing. Science **326:**236-7.

8. **Cheng, H. P., and T. G. Lessie.** 1994. Multiple replicons constituting the genome of Pseudomonas cepacia 17616. J Bacteriol **176:**4034-42.

9. **Coenye, T., and P. Vandamme.** 2003. Diversity and significance of Burkholderia species occupying diverse ecological niches. Environ Microbiol **5:**719-29.

10. **Compant, S., J. Nowak, T. Coenye, C. Clement, and E. Ait Barka.** 2008. Diversity and occurrence of Burkholderia spp. in the natural environment. FEMS Microbiol Rev **32:**607-26.

11. **Curtis, T. P., W. T. Sloan, and J. W. Scannell.** 2002. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A **99:**10494-9.

12. **Dance, D. A.** 2000. Ecology of Burkholderia pseudomallei and the interactions between environmental Burkholderia spp. and human-animal hosts. Acta Trop **74:**159-68.

13. **Dance, D. A.** 1991. Melioidosis: the tip of the iceberg? Clin Microbiol Rev **4:**52-60.

14. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72:**5069-72.

15. **Elliott, G. N., W. M. Chen, J. H. Chou, H. C. Wang, S. Y. Sheu, L. Perin, V. M. Reis, L. Moulin, M. F. Simon, C. Bontemps, J. M. Sutherland, R. Bessi, S. M. de Faria, M. J. Trinick, A. R. Prescott, J. I. Sprent, and E. K. James.** 2007. Burkholderia phymatum is a highly effective nitrogen-fixing symbiont of Mimosa spp. and fixes nitrogen ex planta. New Phytol **173:**168-80.

16. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al.** 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:**496-512.

17. **Gans, J., M. Wolinsky, and J. Dunbar.** 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science **309:**1387-90.

18. **Goris, J., P. De Vos, J. Caballero-Mellado, J. Park, E. Falsen, J. F. Quensen, 3rd, J. M. Tiedje, and P. Vandamme.** 2004. Classification of the biphenyl- and polychlorinated biphenyl-degrading strain LB400T and relatives as Burkholderia xenovorans sp. nov. Int J Syst Evol Microbiol **54:**1677-81.

19. **Hakuno, H., M. Yamamoto, S. Oie, and A. Kamiya.** 2010. Microbial contamination of disinfectants used for intermittent self-catheterization. Jpn J Infect Dis **63:**277-9.

20. **Holden, M. T., H. M. Seth-Smith, L. C. Crossman, M. Sebaihia, S. D. Bentley, A. M. Cerdeno-Tarraga, N. R. Thomson, N. Bason, M. A. Quail, S. Sharp, I. Cherevach, C. Churcher, I. Goodhead, H. Hauser, N. Holroyd, K. Mungall, P. Scott, D. Walker,**

B. White, H. Rose, P. Iversen, D. Mil-Homens, E. P. Rocha, A. M. Fialho, A. Baldwin, C. Dowson, B. G. Barrell, J. R. Govan, P. Vandamme, C. A. Hart, E. Mahenthiralingam, and J. Parkhill. 2009. The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol **191:**261-77.

21. **Holden, M. T., R. W. Titball, S. J. Peacock, A. M. Cerdeno-Tarraga, T. Atkins, L. C. Crossman, T. Pitt, C. Churcher, K. Mungall, S. D. Bentley, M. Sebaihia, N. R. Thomson, N. Bason, I. R. Beacham, K. Brooks, K. A. Brown, N. F. Brown, G. L. Challis, I. Cherevach, T. Chillingworth, A. Cronin, B. Crossett, P. Davis, D. DeShazer, T. Feltwell, A. Fraser, Z. Hance, H. Hauser, S. Holroyd, K. Jagels, K. E. Keith, M. Maddison, S. Moule, C. Price, M. A. Quail, E. Rabbinowitsch, K. Rutherford, M. Sanders, M. Simmonds, S. Songsivilai, K. Stevens, S. Tumapa, M. Vesaratchavest, S. Whitehead, C. Yeats, B. G. Barrell, P. C. Oyston, and J. Parkhill.** 2004. Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proc Natl Acad Sci U S A **101:**14240-5.

22. **Hooper, L. V., and J. I. Gordon.** 2001. Commensal host-bacterial relationships in the gut. Science **292:**1115-8.

23. **Hugenholtz, P., B. M. Goebel, and N. R. Pace.** 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol **180:**4765-74.

24. **Jeong, Y., J. Kim, S. Kim, Y. Kang, T. Nagamatsu, and I. Hwang.** 2003. Toxoflavin produced by Burkholderia glumae causing rice grain rot is responsible for inducing bacterial wilt in many field crops. Plant Dis **87:**890-895.

25. **Kikuchi, Y., X. Y. Meng, and T. Fukatsu.** 2005. Gut symbiotic bacteria of the genus Burkholderia in the broad-headed bugs Riptortus clavatus and Leptocorisa chinensis (Heteroptera: Alydidae). Appl Environ Microbiol **71:**4035-43.

26. **Lackner, G., N. Moebius, L. Partida-Martinez, and C. Hertweck.** 2011. Complete genome sequence of Burkholderia rhizoxinica, an Endosymbiont of Rhizopus microsporus. J Bacteriol **193:**783-4.

27. **Lim, J., T. H. Lee, B. H. Nahm, Y. D. Choi, M. Kim, and I. Hwang.** 2009. Complete genome sequence of Burkholderia glumae BGR1. J Bacteriol **191:**3758-9.

28. **Lipuma, J. J.** 2010. The changing microbial epidemiology in cystic fibrosis. Clin Microbiol Rev **23:**299-323.

29. **Mahenthiralingam, E., A. Baldwin, and C. G. Dowson.** 2008. Burkholderia cepacia complex bacteria: opportunistic pathogens with important natural biology. J Appl Microbiol **104:**1539-51.

30. **Mahenthiralingam, E., T. A. Urban, and J. B. Goldberg.** 2005. The multifarious, multireplicon Burkholderia cepacia complex. Nat Rev Microbiol **3:**144-56.

31. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376-80.

32. **Miche, L., D. Faure, M. Blot, E. Cabanne-Giuli, and J. Balandreau.** 2001. Detection and activity of insertion sequences in environmental strains of Burkholderia. Environ Microbiol **3:**766-73.

33. **Miller, I. M.** 1990. Bacterial leaf nodule symbiosis, p. 163–243. *In* J. A. Callow (ed.), Advances in botanical research, vol. 17.

34. **Moulin, L., A. Munive, B. Dreyfus, and C. Boivin-Masson.** 2001. Nodulation of legumes by members of the beta-subclass of Proteobacteria. Nature **411:**948-50.

35. **Nelson, M. J., S. O. Montgomery, J. O'Neill E, and P. H. Pritchard.** 1986. Aerobic metabolism of trichloroethylene by a bacterial isolate. Appl Environ Microbiol **52:**383-4.

36. **Nierman, W. C., D. DeShazer, H. S. Kim, H. Tettelin, K. E. Nelson, T. Feldblyum, R. L. Ulrich, C. M. Ronning, L. M. Brinkac, S. C. Daugherty, T. D. Davidsen, R. T. Deboy, G. Dimitrov, R. J. Dodson, A. S. Durkin, M. L. Gwinn, D. H. Haft, H. Khouri, J. F. Kolonay, R. Madupu, Y. Mohammoud, W. C. Nelson, D. Radune, C. M. Romero, S. Sarria, J. Selengut, C. Shamblin, S. A. Sullivan, O. White, Y. Yu, N. Zafar, L. Zhou, and C. M. Fraser.** 2004. Structural flexibility in the Burkholderia mallei genome. Proc Natl Acad Sci U S A **101:**14246-51.

37. **O'Sullivan, L. A., and E. Mahenthiralingam.** 2005. Biotechnological potential within the genus Burkholderia. Lett Appl Microbiol **41:**8-11.

38. **Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen.** 1985. Analyzing natural microbial populations by rRNA sequences, p. 4-12, ASM News, vol. 51.

39. **Partida-Martinez, L. P., and C. Hertweck.** 2005. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. Nature **437:**884-8.

40. **Roesch, L. F., R. R. Fulthorpe, A. Riva, G. Casella, A. K. Hadwin, A. D. Kent, S. H. Daroub, F. A. Camargo, W. G. Farmerie, and E. W. Triplett.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J **1:**283-90.

41. **Santos, A. V., R. J. Dillon, V. M. Dillon, S. E. Reynolds, and R. I. Samuels.** 2004. Ocurrence of the antibiotic producing bacterium Burkholderia sp. in colonies of the leaf-cutting ant Atta sexdens rubropilosa. FEMS Microbiol Lett **239:**319-23.

42. **Schmidt, T. M., E. F. DeLong, and N. R. Pace.** 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol **173:**4371-8.

43. **Staley, J. T., R. W. Castenholz, R. R. Colwell, J. G. Holt, M. D. Kane, N. R. Pace, A. A. Salyers, and J. M. Tiedje.** 1997. The Microbial World: Foundation of the Biosphere. American Academy of Microbiology.

44. **Talbi, C., M. J. Delgado, L. Girard, A. Ramirez-Trujillo, J. Caballero-Mellado, and E. J. Bedmar.** 2010. Burkholderia phymatum strains capable of nodulating Phaseolus vulgaris are present in Moroccan soils. Appl Environ Microbiol **76:**4587-91.

45. **Torsvik, V., J. Goksoyr, and F. L. Daae.** 1990. High diversity in DNA of soil bacteria. Appl Environ Microbiol **56:**782-7.

46. **Van Oevelen, S., R. De Wachter, P. Vandamme, E. Robbrecht, and E. Prinsen.** 2004. 'Candidatus Burkholderia calva' and 'Candidatus Burkholderia nigropunctata' as leaf gall endosymbionts of African Psychotria. Int J Syst Evol Microbiol **54:**2237-9.

47. **Van Oevelen, S., R. De Wachter, P. Vandamme, E. Robbrecht, and E. Prinsen.** 2002. Identification of the bacterial endosymbionts in leaf galls of Psychotria (Rubiaceae, angiosperms) and proposal of 'Candidatus Burkholderia kirkii' sp. nov. Int J Syst Evol Microbiol **52:**2023-7.

48. **Vandamme, P., J. Goris, W. M. Chen, P. de Vos, and A. Willems.** 2002. Burkholderia tuberum sp. nov. and Burkholderia phymatum sp. nov., nodulate the roots of tropical legumes. Syst Appl Microbiol **25:**507-12.

49. **Vial, L., M. C. Groleau, V. Dekimpe, and E. Deziel.** 2007. Burkholderia diversity and versatility: an inventory of the extracellular products. J Microbiol Biotechnol **17:**1407-29.

50. **Wigley, P., and N. F. Burton.** 2000. Multiple chromosomes in Burkholderia cepacia and B. gladioli and their distribution in clinical and environmental strains of B. cepacia. J Appl Microbiol **88:**914-8.

51. **Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk, and J. A. Eisen.** 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature **462:**1056-60.

52. **Yoder-Himes, D. R., P. S. Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, J. M. Tiedje, and R. Sorek.** 2009. Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. Proc Natl Acad Sci U S A **106:**3976-81.

53. **Zhao, N., C. Qu, E. Wang, and W. Chen.** 1995. Phylogenetic evidence for the transfer of Pseudomonas cocovenenans (van Damme et al. 1960) to the genus Burkholderia as Burkholderia cocovenenans (van Damme et al. 1960) comb. nov. Int J Syst Bacteriol **45:**600-3.

Chapter 2.
Genomes and genomic diversity within the *Burkholderia* genus.

Introduction

The *Burkholderia* genus is comprised of an extremely vast group of Betaproteobacteria whose phenotypic diversity and ability to thrive in varied ecological settings are equaled by their genomic heterogeneity. The first *Burkholderia* isolates were described in the 1940s by Walter H. Burkholder (3), though at the time, phenotypic observations led these organisms to be misclassified as Pseudomonads. DNA-DNA hybridization experiments conducted in the 1970s revealed little homology among many bacteria categorized as Pseudomonads and further refinement of this group based on rRNA genetic similarity spawned the creation of the new *Burkholderia* genus in its own family/class (50). Now, with 65 formally accepted *Burkholderia* species named and many others being characterized, this group represents one of the most ecologically diverse clades known. These organisms are ubiquitous in the environment, living in soil, water, within plant, insect, and animal hosts, are contaminants of industrial materials and processes, and exhibit lifestyles that run the gamut from those of beneficial symbionts to malignant pathogens of crops, animals and humans (see Chapter 1 for overview).

It has been suggested that the great capacity of this genus to adapt to a wide spectrum of niches is owed to its metabolic versatility by virtue of a large and highly dynamic genome (7, 25). The study of their genomes, together with their phenotypes, distribution in the environment and their evolutionary relationships have begun to shed light on the potential for some *Burkholderia* to be exploited as agents of bioremediation and crop management, as well as providing a better understanding of how others threaten animal and crop health. The first genomic studies of this group were targeted toward *B. pseudomallei* (20) and *B. mallei* (30), due in part to their historical use as biological

weapons and their subsequent characterization by the Centers for Disease Control and prevention as Category B bioterrorism agents. These and follow-on studies revealed that these genomes are highly plastic, with many genomic islands that are variably present within the species, and have undergone a number of genomic rearrangements and deletions (24). These two highly related genomes were followed by a description of the very different and versatile biodegradative *B. xenovorans* strain LB400 (7), which harbors a large megaplasmid that appears to be responsible for many of the interesting capabilities in degrading aromatic contaminants, in addition to the two chromosomes conserved among most *Burkholderia* spp. Interestingly, we noted that the megaplasmid and secondary chromosome appeared to be under relaxed selective pressure and are much less conserved than the main chromosome. Few studies have looked at this feature in more detail, though it has been suggested that codon usage may have played a role in this evolutionary process (11, 28).

Due to the many recent advances in sequencing, and in methods to explore and analyze the resulting data, many other genome projects have been initiated, including some targeted for genome closure, and others only to draft for comparison with completed reference genomes. These projects have already begun to allow the exploration of the genetic underpinnings of phenotypic characteristics via comparative genomic efforts, as well as enable the exploration of novel methods to interrogate *Burkholderia* evolution, not only at the gene level but on the whole genome level, also termed phylogenomics. Although a number of efforts have either used microarray or other data to explore such questions (39), or have used comparative genomics for the detection of specific species (19), there

have been limited studies that try to compare available sequence data for *Burkholderia*, either the entire genus or particular species (28, 45).

Here, we characterize the genus of *Burkholderia* by undertaking an in-depth comparative genomics approach, analyzing all 25 available completed genomes, representing 13 formally accepted species and an additional 2 representatives of unnamed species. Within this set are included three novel lineages, two of which lie within a large underexplored group of *Burkholderia* known primarily for their beneficial association with plants. The available genomes include four genomes each of *B. mallei*, the obligate pathogen of equines, and *B. pseudomallei*, a tropical soil organism and causative agent of mellioidosis in humans and animals, along with the closely related *B. thailandensis* which is often considered non-virulent, being only rarely associated with human disease (17).

There are also nine completed genomes of the *Burkholderia cepacia* complex (*Bcc*), who are more distantly related to the Pseudomallei group, but still of public health concern, particularly within the cystic fibrosis community (and patients with chronic granulomatous disease) because of their ability to be transmitted from patient to patient, their capacity to cause necrotising invasive infection and death, and their natural resistance to treatment therapies, including antibiotics (25). However, not all species within this closely related group are equally transmissible (41), nor are they equally distributed in the environment, implying that differences within the genome should be discernable. Within the *Bcc*, there are four *B. cenocepacia* genomes available (two clinical specimens and two isolated from soil), two rhizosphere *B. ambifaria* genomes, the type strain of *B. multivorans* (originally isolated from soil), *B. vietnamiensis* strain G4, the best known trichloroethene oxidizer

isolated from an industrial waste treatment plant, and the forest soil isolate *B. lata* strain 383.

In addition to the *Bcc* and Pseudomallei group, a genome is complete for *B. glumae*, a rice pathogen that has also been recovered in a patient with chronic granulomatous disease (48) but whose phylogenetic position within the genus has been ambiguous with respect to these other two pathogenic groups. The recently sequenced genome of the endosymbiont *B. rhizoxinica* is also available (23). This organism is of particular interest due to its close mutualistic symbiosis with fungi, where it has been shown to supply the fungal host with rhizoxin, an antimitotic virulence factor, thus contributing to rice seedling blight. *B. rhizoxinica* has also been shown to be transmitted vertically with its host and interestingly, has been recently associated with human clinical specimens (16).

Finally, there are a number of other completed genomes that belong to a large group of *Burkholderia* best known for their biodegradation capabilities, such as *B. xenovorans* LB400, and for their beneficial symbiotic associations with plants. The latter group includes *B. phymatum*, the first example of a betaproteobacterium capable of symbiotic nodulation (27), and the first reported beta-rhizobial strain to fix nitrogen in free-living culture (14), as well as *B. phytofirmans*, an endophytic plant growth-promoting rhizobacterium (38). In addition to these, are two novel and as yet unnamed species (sp. CCGE1002 and sp. CCGE1003) isolated from grasslands and phylogenetically most similar to *B. graminis*, a known rhizosphere colonizer (of wheat in the case of the sequenced type strain).

Given the breadth of the diversity of phenotypes represented within this short list of organisms, all belonging to the same genus, the *Burkholderia*, we wished to clarify the

phylogenetic relationships within the genus, to better understand - within an evolutionary framework - the genomic variation between species, to uncover both the common and unique gene sets of the genus and of the various clades within the genus, and to broadly relate observed differences to their various phenotypes, lifestyles and niche preferences.

Materials and Methods

Genomes analyzed

Due to the detailed nature of some of the comparisons we planned to undertake, we took all precautions to avoid analyzing sequencing errors, and thus excluded all incomplete genomes from our analyses (>60 - see http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi for up-to-date list). All 25 available completed *Burkholderia* genomes were obtained from Genbank (February 2011) and are listed in Table 2.1.

Whole genome alignments and visualization

Genomes in the form of fasta sequences from selected representative species were concatenated in order of chromosome, and then by plasmid name. The program TblastX was used to perform local alignments between selected pairs of genomes, and later filtered by E value, bit score and size of hit, with precise parameters that varied based on the genetic relatedness of each pairwise comparison (1). The visualization program ACT was used to display each pairwise alignments (4).

Table 2.1. Sequenced and completed *Burkholderia* genomes.

Genome	Size (Mb)	G+C	Genome Center* (Year)	Description
***B. ambifaria* AMMD**	7.52	66.8	JGI (2006)	*Bcc*, soil isolate
***B. ambifaria* MC40-6**	7.64	66.4	JGI (2008)	*Bcc*, soil isolate
***B. cenocepacia* AU 1054**	7.28	66.9	JGI (2006)	*Bcc*, clinical isolate
***B. cenocepacia* HI2424**	7.72	66.8	JGI (2006)	*Bcc*, soil isolate
***B. cenocepacia* J2315**	8.04	66.9	WTSI (2008)	*Bcc*, clinical isolate
***B. cenocepacia* MC0-3**	7.96	66.6	JGI (2008)	*Bcc*, soil isolate
***B. glumae* BGR1**	7.27	67.9	SNU (2009)	Rice pathogen
***B. lata* 383**	8.69	66.3	JGI (2004)	*Bcc*, forest soil
***B. mallei* ATCC 23344**	5.83	68.5	TIGR (2004)	Obligate pathogen
***B. mallei* NCTC 10229**	5.76	68.5	TIGR (2007)	Obligate pathogen
***B. mallei* NCTC 10247**	5.85	68.5	TIGR (2007)	Obligate pathogen
***B. mallei* SAVP1**	5.23	68.4	TIGR (2007)	Avirulent *B. mallei*
***B. multivorans* ATCC 17616**	7.01	66.7	JGI (2007)	*Bcc*, soil isolate
***B. phymatum* STM815**	8.68	62.3	JGI (2008)	Nodulator, N2 fixer
***B. phytofirmans* PsJN**	8.21	62.3	JGI (2008)	Endophytic symbiont
***B. pseudomallei* 1106a**	7.10	68.3	TIGR (2007)	Clinical pathogen
***B. pseudomallei* 1710b**	7.30	68.0	TIGR (2005)	Clinical pathogen
***B. pseudomallei* 668**	7.03	68.3	TIGR (2007)	Clinical pathogen
***B. pseudomallei* K96243**	7.25	68.1	WTSI (2004)	Clinical pathogen
***B. rhizoxinica* HKI 454**	3.75	60.5	HKI (2011)	Fungal endosymbiont
***Burkholderia* sp. CCGE1002**	7.89	63.3	JGI (2009)	Grassland rhizosphere
***Burkholderia* sp. CCGE1003**	7.10	63.3	JGI (2010)	Grassland rhizosphere
***B. thailandensis* E264**	6.71	67.6	TIGR (2005)	Soil saprophyte
***B. vietnamiensis* G4**	8.39	65.7	JGI (2007)	*Bcc*, wastewater, TCE degrader
***B. xenovorans* LB400**	9.74	62.6	JGI (2006)	Landfill, PCB degrader

*JGI: DOE Joint Genome Institute; WTSI: Wellcome Trust Sanger Institute; TIGR: The Institute for Genomic Research; SNU: Seoul National University; HKI: Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute

Ortholog identification

Orthologs amongst the *Burkholderia* genomes were identified by the program Ortholuge (15) using *Ralstonia solanacearum* GMI1000 as the outgroup. Briefly, the reciprocal-best-blast hits were determined for all genes in each genome using a many-to-many blastP search (1) with an expectation value (E-value) cutoff of 1e-4. Each candidate was then aligned with the others using ClustalW (44) and their phylogenetic distance ratios calculated with fprotdist, PHYLIP's protdist software compiled with EMBOSS as an EMBASSY package. Orthologs between genomes were then identified using ratio cutoff values of R1 ≤ 0.55 and R2 ≤ 0.70, as recommended. Orthologs not shared with the outgroup were detected using a reciprocal best blast solution. BLAST hits for each gene were collected if the top BLAST hit (cutoff of 1e-20) and at least 70% of the sequence in the alignment and with 60% or greater percent identity. If the top BLAST hit for a gene in genome A is the top BLAST hit of a gene in genome B, then a paired match is formed and these genes are designated orthologs.

Pangenome analysis

The *Burkholderia* pan genome was created in two major stages. The first stage involved the identification of ortholog families using the pairwise relationships identified by Ortholuge and merging those families that had broken links between each other. The second stage involved the integration of orthologs identified by the reciprocal best blast hit method into the growing ortholog families. Six tests were performed to assess the validity of these ortholog families representing the pan genome. First, all orthologs identified in Ortholuge must remain in families of no less than 2 members after the merging process

(pairwise limit). Second, no ortholog family can have more ortholog sequences than there are genomes involved in the analysis (family gene limit). Third, no single gene can be a member in more than one family (non-redundant gene). Fourth, no ortholog family may consist of multiple genes from a single organism (non-redundant source). Fifth, all orthologs identified via Ortholuge or the reciprocal best blast hit method must remain in the pan genome after any and all family merging cycles (gene persistence). Sixth, all genes from within the organism's FASTA files must be placed uniquely into the pangenome (equal gene representation). COG classifications (43) were assigned to each gene, if available, by parsing the PID from the *.ptt file that resides in the genome project FTP site at NCBI. Lookup tables were created to link the NC numbered files with the GI, COG, chromosome number, and sequence information (obtained by the *.faa files).

Phylogenetic analyses

For conserved 'housekeeping' protein trees, protein marker sequences for 31 proteins were retrieved from each *Burkholderia* genomes and also that of the outgroup *Ralstonia solanacearum* using the MarkerScanner.pl utility script from AMPHORA (49), which uses profile hidden Markov models to search for these conserved proteins. The proteins used are encoded by conserved housekeeping genes involved in information processing or central metabolism and likely not involved in frequent lateral gene transfer (*dnaG*, *ffr*, *infC*, *nusA*, *pgk*, *pyrG*, *rplABCDEFKLMNPST*, *rpmA*, *proB*, *rpsBCEIJKMS*, *smpB*, *tsf*). Although a few of the genes were not located within the annotations of some of the genomes, similarity searches revealed that these were simply annotation errors, and the genes were added to the list. The 31 protein sequences were concatenated by species and

used to perform a multiple sequence alignment using MUSCLE v3.8.31 (13). Positions for which the assignment of homology is uncertain were excluded from further analysis by masking and Gblocks (version 0.91b) was used to select conserved blocks for phylogenetic analysis (42) with adjusted parameters (Minimum Number Of Sequences For A Conserved Position: 14, Minimum Number Of Sequences For A Flank Position: 14, Maximum Number Of Contiguous Nonconserved Positions: 15, Minimum Length Of A Block: 3, and Allowed Gap Positions: 50%).

For core genome trees, the protein sequences of the core genome (as calculated via pangenomic ortholog analysis) were selected and concatenated together. A multiple sequence alignment was performed using MAFFT (22), and the conserved blocks were selected using GBLOCK as described above. A maximum likelihood phylogenetic tree was constructed from the masked concatenated protein alignment of 31 housekeeping genes using PHYML v3.0.1 (18), and an approximately maximum likelihood phylogeny was constructed for the core protein sequences using FastTree2 (37). One hundred bootstrapped replicates were applied to each phylogeny. The best model, selected based on a likelihood ratio test, was the Jones-Taylor-Thorton (JTT) model of amino acid substitution. The program FigTree (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize and decorate the tree.

Results

Inferring phylogeny of sequenced members of the *Burkholderia* genus

Given that a phylogeny of all the species with completed genomes has not yet been presented, we built such a phylogeny and also tried to resolve a number of incongruities observed by different groups. First, the placement of *B. glumae* has been shown either: 1) placed within a lineage shared with the *Bc*c but exclusive of the Pseudomallei group (e.g. (10, 16, 34)), typically using the 16S rRNA sequence; 2) as an outgroup to both the Pseudomallei and *Bc*c lineages (40) using multiple locus sequence typing (MLST) analysis; or 3) even within the *Bc*c (5), using 16S sequence. Second, the recently sequenced *B. rhizoxinica* has been reported in one 16S study as a sister clade to the *B. graminis* group, with *B. glathei* and the Pseudomallei and *Bc*c clades as outgroups (16), which conflicts with another 16S report that it is an outgroup to all of these species, including *B. graminis* (5).

Here, we constructed a phylogenomic tree inferred from the concatenated alignment of 31 conserved housekeeping proteins shared among the *Burkholderia* and *Ralstonia solanacearum* (Figure 2.1). All species relationships within the three major clades corroborate the topologies of previous efforts (highlighted in green, purple and red). It is clear that with these 31 conserved genes, the *B. mallei* and even some of the *B. pseudomallei* sequences are too similar to be able to discriminate among them adequately. Of note, *B. glumae* is definitively placed outside the *Bc*c and Pseudomallei group with high bootstrap support in this phylogeny, in contrast to 16S rRNA-derived trees, and interestingly the new *B. rhizoxinica* genome is placed as a significantly distant outgroup to the other completed *Burkholderia* genomes.

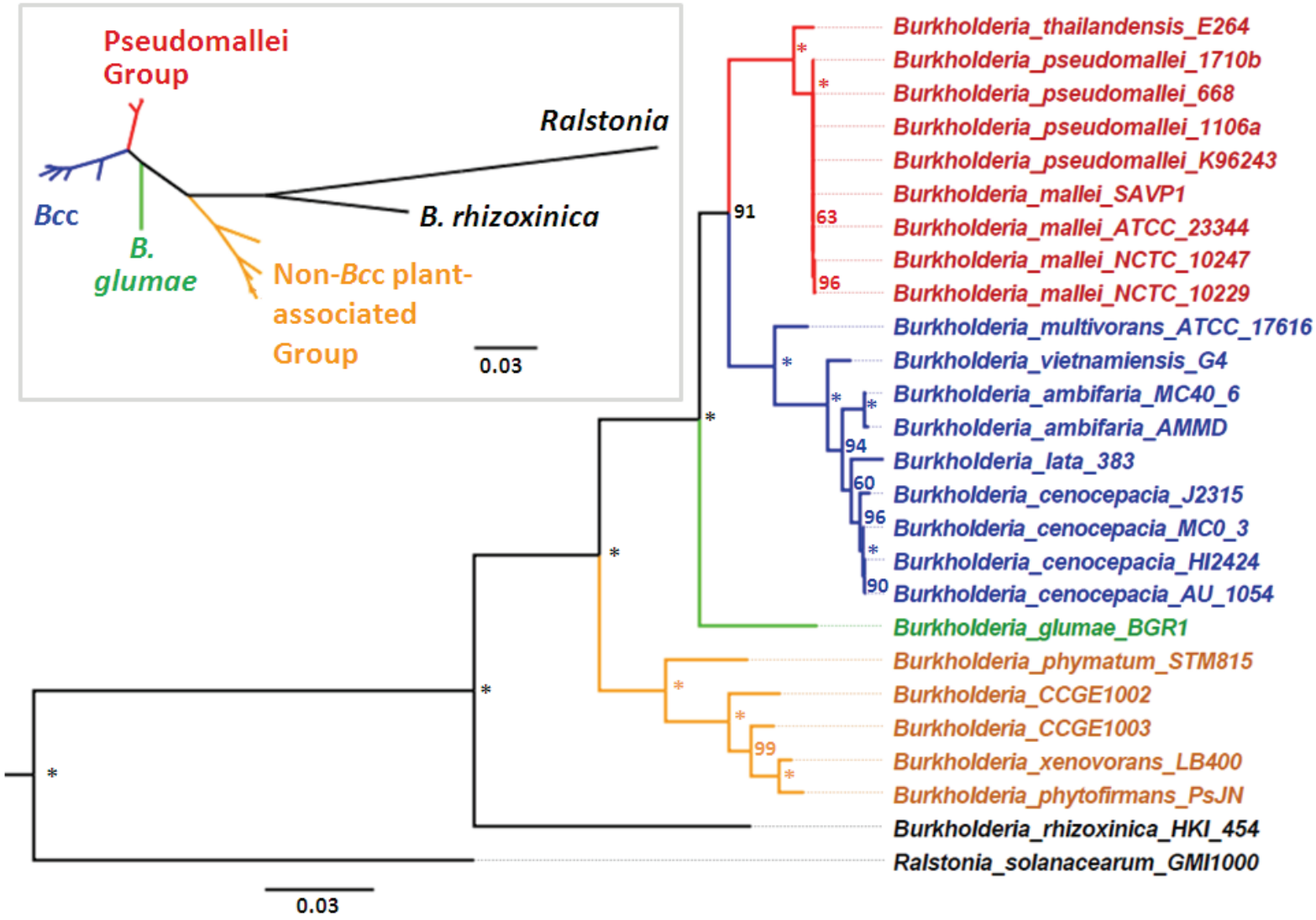


Figure 2.1. *Burkholderia* genus phylogeny. A 31-gene phylogram of all *Burkholderia* species used in this study is presented with *Ralstonia* as an outgroup. Bootstrap confidence values indicated above 50% (* = 100%). Inset is the same tree, unrooted.

Whole genome alignments and genome plasticity

Given the phylogeny presented in Figure 2.1, and observations regarding differential chromosome evolution(7), the occurrence of different genomic islands (20), and many rearrangements among *B. mallei* strains (30), we sought to compare and visualize the genome structure and genome synteny of twelve representative lineages within the *Burkholderia* by performing whole genome alignments (Figure 2.2). While it is clear that the *Burkholderia* genome is rather fluid, with many rearrangements apparent, the majority of chromosome 1 appears to be well conserved throughout the *Burkholderia* spp. Although many syntenic regions in chromosome 2 can also be discerned from these alignments, particularly among more closely related species (see Figure 2.1 for phylogenetic distance), it is readily apparent that genome synteny (or similarity in general) deteriorates as a function of phylogenetic distance (e.g. *B. rhizoxinica* vs *B. phytofirmans*, or *B. phymatum* vs *B. glumae*). Furthermore, while the *Bc*c (shaded blue in Figure 2.1) all share a third chromosome, this replicon is only found within this clade and appears to evolve rapidly when compared with the other two chromosomes. Each species has also either acquired novel material not present in its nearest relatives (e.g. genomic islands acquired via lateral gene transfer), or undergone large deletion events, shown as regions absent from the genome yet present in others. Due to the large amount of diversity that exists among all genomes, subsequently only protein-based comparisons were performed since rigorous nucleotide-based analyses would not be possible.

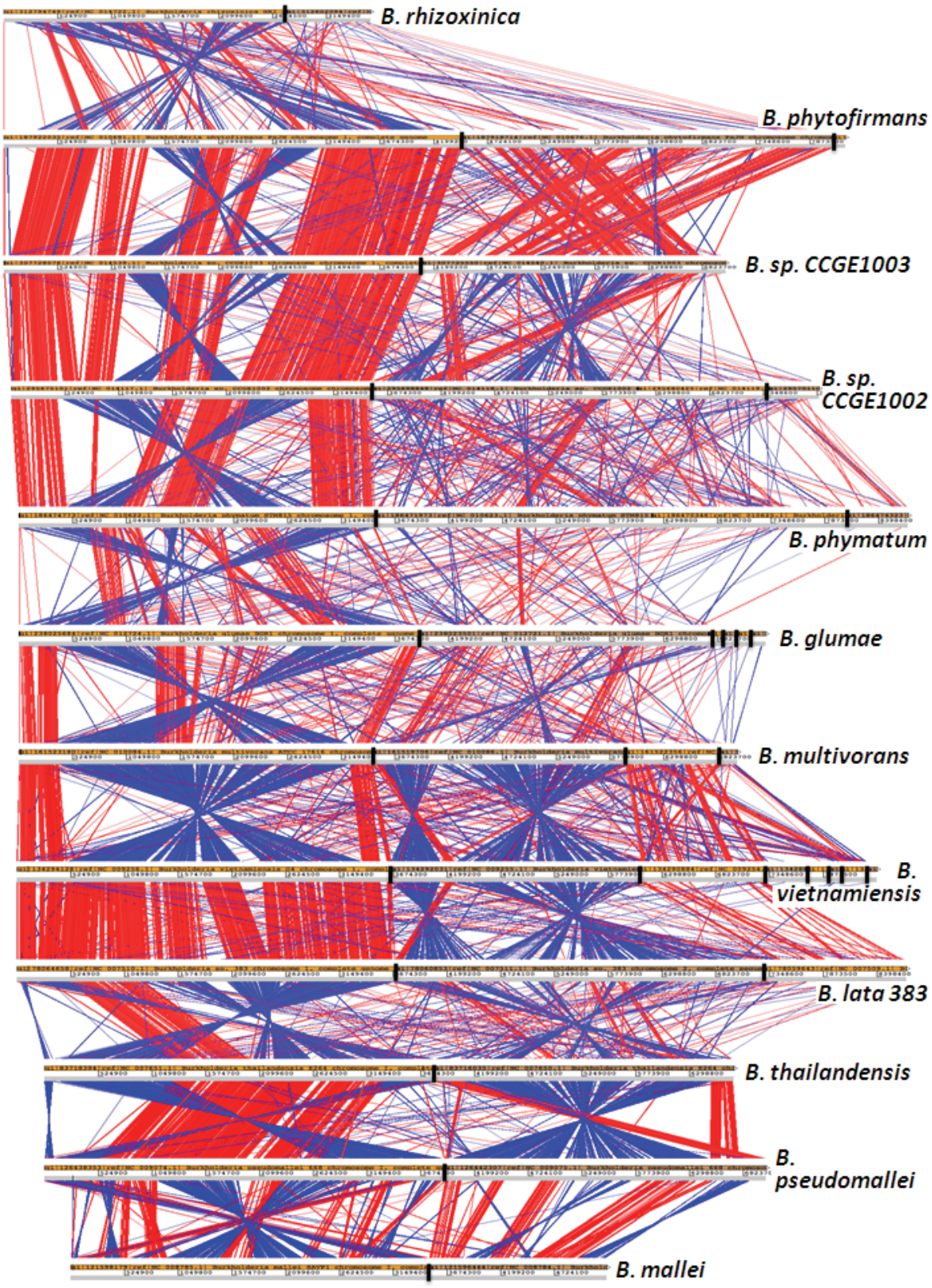


Figure 2.2. Whole genome alignments of *Burkholderia* genomes. The genomes of

twelve representative *Burkholderia* species (only representative strains 668 and SAVP1 were selected for *B. pseudomallei* and *B. mallei*, respectively, and only 3 members of the closely related *Bcc* were chosen, *B. multivorans*, *B. vietnamiensis* and *B. lata*) were concatenated, such that all chromosomes and plasmids were placed end to end. The chromosomes and plasmids are delineated by black lines and represented as orange or brown horizontal boxes in the order: chromosome 1, chromosome 2, etc. Synteny in the form of red blocks of parallel colinearity, and blue blocks of inverted colinearity are shown between genomes.

Burkholderia genus pangenome

While whole genome alignments can be made to observe the progressive differences among species that are less and less phylogenetically similar, an alternate method is required for a more comprehensive genus-wide comparison of such distinctly varied and divergent genomes. We have undertaken a pangenome approach, where the encoded proteins from each of the available genomes were compared to each other, as well as an outgroup (the genome of *R. solanacearum*), in order to assess orthology. This method allows the estimation of the core orthologous genes (shared) within the entire *Burkholderia* genus, the variable (also called auxiliary, adaptive, character, etc.) gene families present in two or more of the genomes, and the "unique genome" for all species and strains examined (Figure 2.3). While the average genome is <7Mb in size with <7000 genes, the pangenome of these 25 *Burkholderia* representatives is more than eight times as large and totals 56,777 gene families. The *Burkholderia* core is but 829 genes, or ~12.5% of the typical

Burkholderia genome, the variable genome encompasses 12,701 gene families shared among two or more but not all genomes, and the unique genome includes an amazing 43,247 gene families, or >75% of the *Burkholderia* pangenome (Figure 2.3). Despite this large figure, the average unique portion of genes within *Burkholderia* genomes is roughly 26.7%, while the 829 gene core accounts for 13.3% on average and the remaining >50% consists of variable genes (Figure 2.3).

This pangenome diversity is essentially an underestimate, since four species are represented by 14 of the 25 genomes, which results in significant gene family redundancy and an underestimate of the expected unique and variable fractions (and potentially an overestimate of the core) if 25 randomly selected different species were available and used for this analysis. Indeed when only one member of each species was used to calculate the pangenome (n=15, one genome per species), the core increases minimally, as expected, to 869, while the total pangenome size is moderately reduced to 42,485 (averaging a contribution of more than 2,830 gene families per genome, compared with 2,250 when all 25 genomes are included). Undoubtedly, bias and non-randomness in the available completed genome sequences will always contribute to such underestimates of pangenome size.

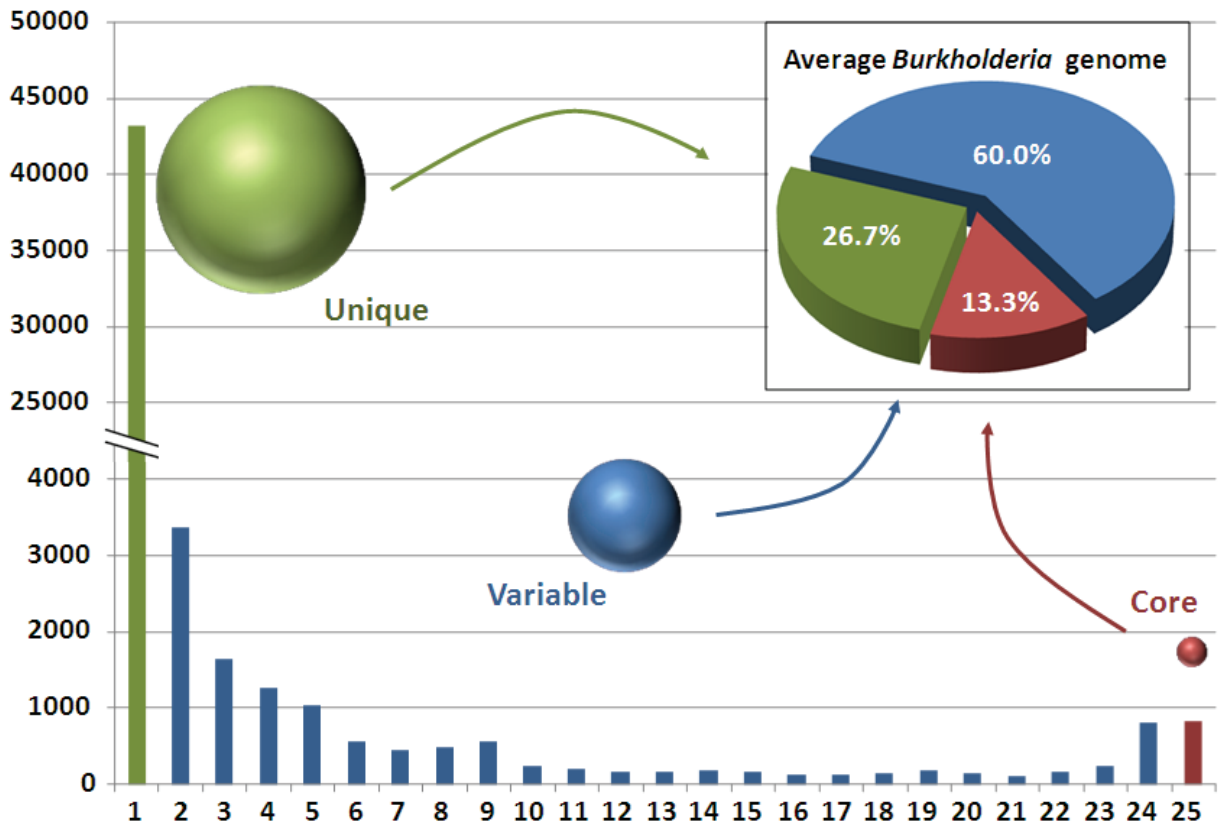


Figure 2.3. The *Burkholderia* pangenome core, variable and unique genome. The histogram displays the number of gene families (Y axis) that can be found in X number of genomes (X axis), thus green represents those gene families found only once in a single genome, red indicates those found in all the genomes, and blue corresponds to genes found in some (2 or more) but not all genomes. The size of the balls represents the relative total number of protein families present in these three groups: 43,247 Unique genes, 12,701 Variable genes, 829 Core genes. Inset is the average *Burkholderia* genome composition, displaying the contribution of Core, Variable and Unique genes. Thus, despite the far larger pool of unique genes, this category represents roughly only 26.7% of the average genome.

In addition, the genome of *B. rhizoxinica* dramatically alters the pangenome structure, primarily due to its small size. At 3.75 Mb, it is less than half the average genome size with the exception of the *B. mallei* group which themselves are undergoing genome reduction (~5.5Mb in size). When *B. rhizoxinica* is removed from the pangenome analysis, the core genome size almost doubles to 1,506, while minimally impacting the total pangenome size (54,693 gene families). This is apparent in the distribution of gene families among genomes (Figure 2.3), where >800 gene families are present in 24 of the 25 genomes, most of which lack a *B. rhizoxinica* ortholog.

When the core, variable and unique gene families are broken down by the replicon on which they reside, an interesting pattern emerges. Due to the fact that a number of genomes carry two chromosomes while others carry three, the genes were distinguished based on whether they were consistently present on the main chromosome, consistently present on secondary replicons, or found on the main chromosome in some, but not all the genomes (Figure 2.4). The majority of core genes are always found on the main chromosome, and while a number are also occasionally found on secondary replicons, a very small fraction is always found on secondary replicons. This is consistent with previous findings. The variable genes are true to their name, and do not carry a clear signal, with large fractions of these genes found throughout the genome (Figure 2.4). Interestingly almost 40% of the unique genes are also found on the main chromosome, and while one might expect a larger fraction of unique genes to originate from secondary replicons, the main chromosome is often substantially larger than the other replicons. As may be expected given their distribution among replicons, in terms of functional distribution, the core differs from the variable and unique genes in the proportion of genes

involved in translation, post-translation modification, nucleotide metabolism and cell division and cell envelope biogenesis. Meanwhile, the variable and unique genomes are overrepresented in signal transduction, transcriptional regulation, secondary metabolism, and sugar and amino acid transport and metabolism.

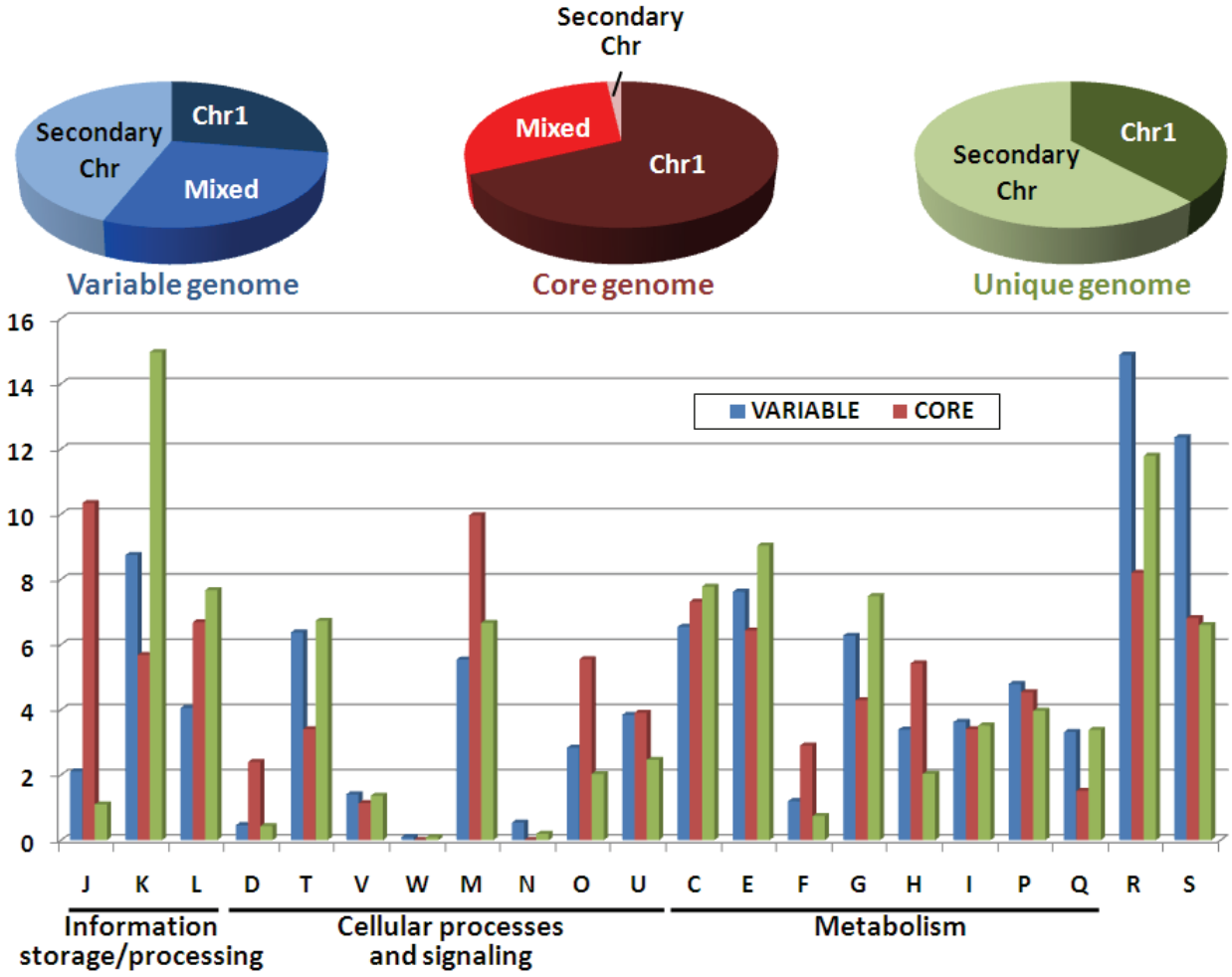


Figure 2.4. Replicon and functional distribution of core, variable and unique gene families. The distribution of the core (red), variable (blue) and unique (green) genes with respect to chromosome (chromosome 1 or other; piecharts above), and with respect to COG functional category (graph below). The COG categories are as follows, J: Translation,

ribosomal structure and biogenesis; K: Transcription; L: Replication, recombination and repair; D: Cell cycle control, cell division, chromosome partitioning; T: Signal transduction mechanisms; V: Defense mechanisms; W: Extracellular structures; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Posttranslational modification, protein turnover, chaperones; U: Intracellular trafficking, secretion, and vesicular transport; C: Energy production and conversion; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R: General function prediction only; S: Unknown.

Comparing the pangenomes of distinct lineages of *Burkholderia*

A phylogeny was inferred from the 829 concatenated core genes and the resulting tree maintained the same gross topology, although it differed at a few positions, namely placing *B. ambifaria* genomes along with *B. vietnamiensis* as sister taxa, and differences in the poorly resolved *B. pseudomallei*/*B. mallei* strains (not shown). From both the 31 housekeeping gene phylogeny and that inferred from the core *Burkholderia* genome however, it is clear that the genus is divided into several independently clustering groups of species (Figure 2.1): the Pseudomallei group, the *Bc*c species, and the "non-*Bc*c plant-associated" group. In addition, upon examination of gene families from the pangenome analysis, it is also clear that there are several blocks of genes that are only commonly shared among one of the three groups (Figure 2.5). We therefore sought to examine the

pangenomes of these clades and understand what protein families were present only within one of the lineages, and which ones were only sporadically shared with some of the other clades.

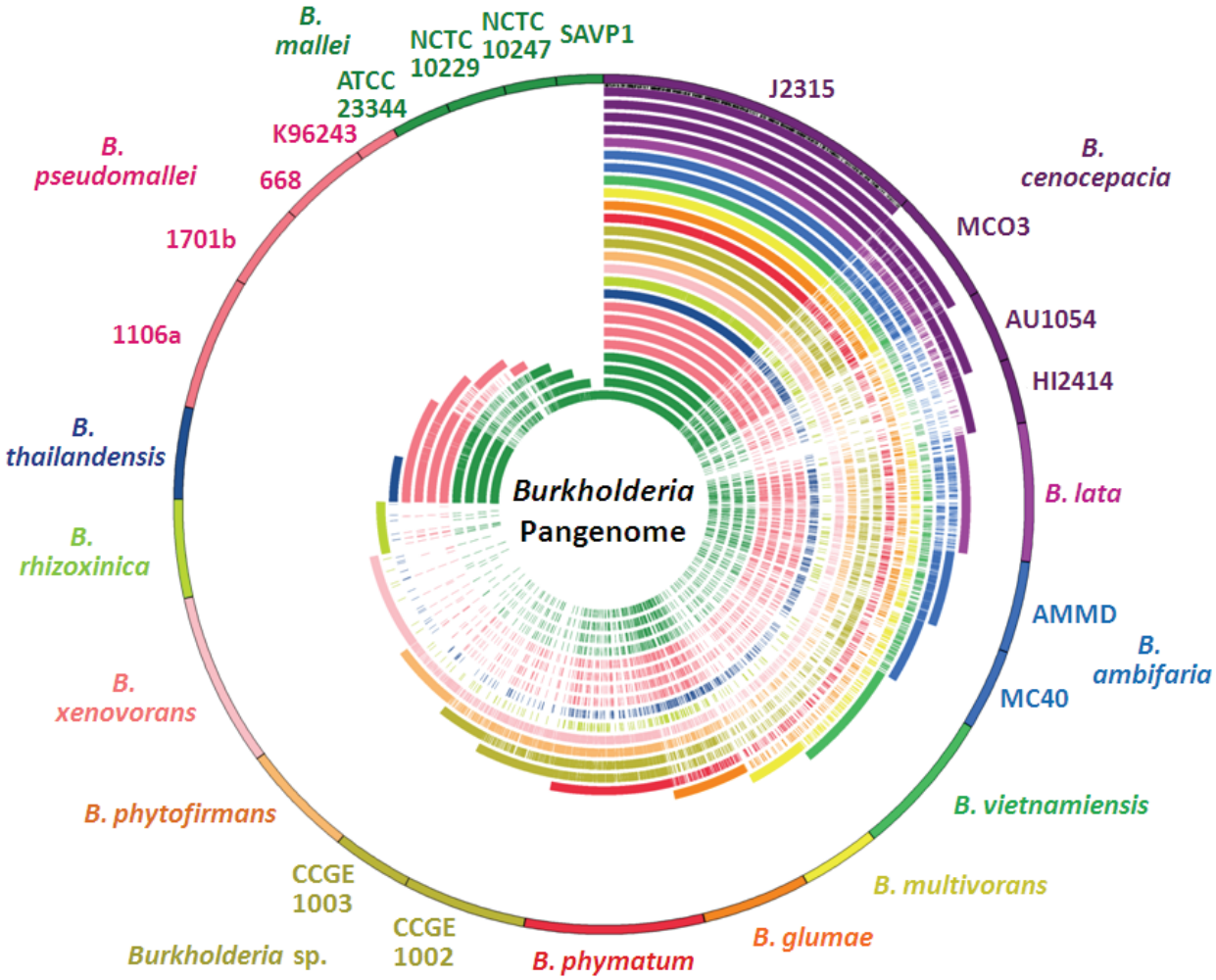


Figure 2.5. Pangenome of the *Burkholderia* genus. Comparative gene content of 25 genomes of *Burkholderia*. Beginning clockwise from the top of the circle, the outermost circle represents the 56,777 gene family *Burkholderia* pangenome, all genes (with constant wedge width) present originating from the genome whose name is indicated outside the circle. Only the genes not present (ie. that do not have orthologs) in any of the preceding

genomes are displayed (outer circle), such that each wedge indicates a unique gene (protein) family. For all 25 inner circles, all genes are shown that match the gene family represented in the outer circle. The first genome is that of *B. cenocepacia* J2315 and proceeds to *B. mallei* SAVP1. This order is the same clockwise as it is for outermost to innermost circles, and they are also colored accordingly. Genomes of different strains of the same species are colored the same, and genes are ordered with respect to position in the chromosomes in chromosome order. The small black lines between the outer circle and the J2315 ring are the core genes shared among all genomes.

Similar to the *Burkholderia* genus pangenome, the pangenomes for the three well sequenced lineages depicted in Figure 2.1 were calculated. The core genomes were calculated for all three lineages, the *Bcc*, the Pseudomallei group (PseudoGp), and the non-*Bcc* rhizosphere isolates (Rhizo) (Figure 2.6). The lineage-specific genes were further screened for presence within the larger pangenome of the other two lineages to identify less conserved orthologs within those groups. The depleted set of lineage-specific genes was analyzed for functional category and for residence on either the main chromosome or one of the secondary chromosomes (Figure 2.6).

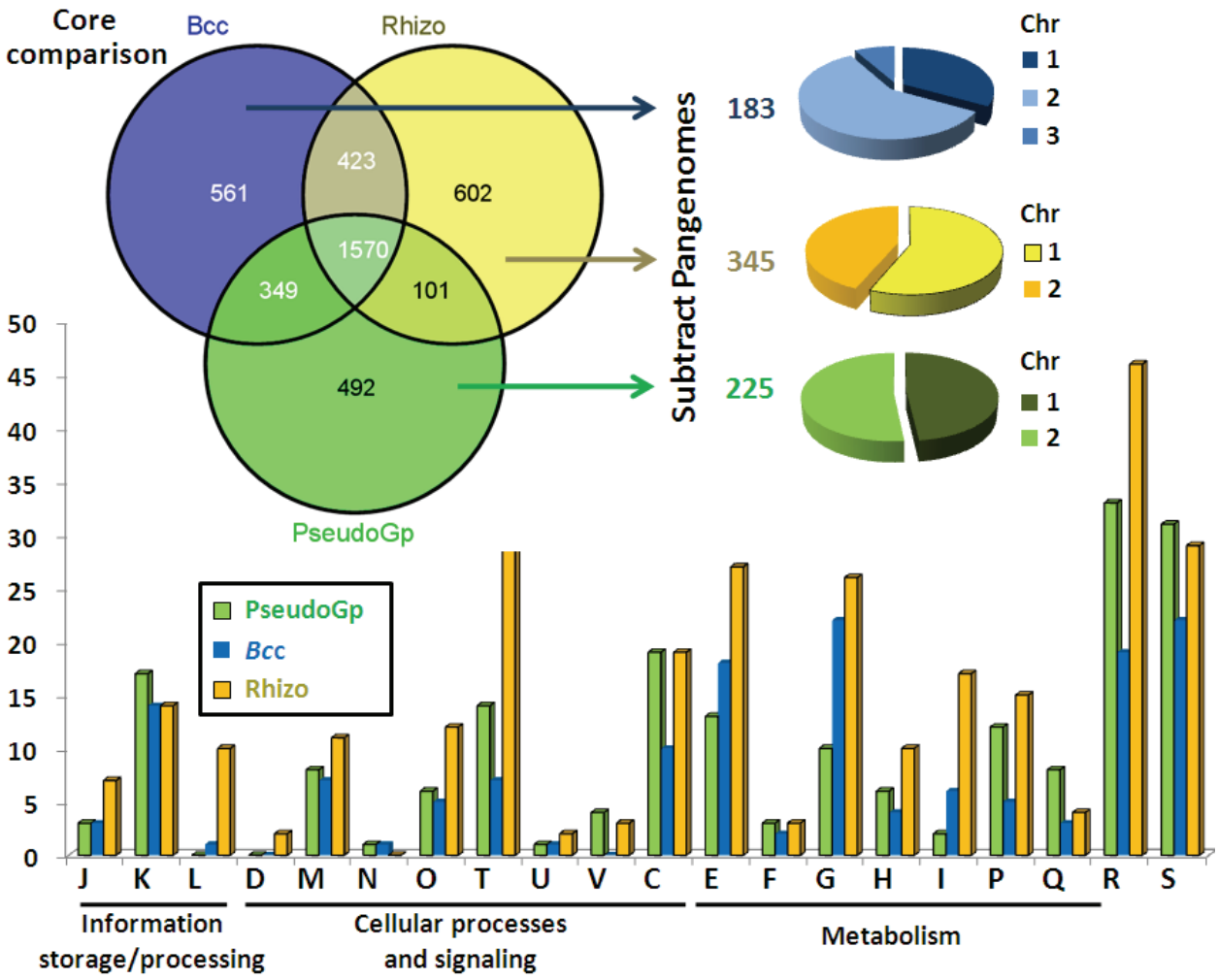


Figure 2.6. Lineage-pangenome comparisons. Above is a Venn diagram depicting the core genomes of the 3 main lineages of *Burkholderia*, their overlap, genomic location and functional capacity. The non-core pangenomes were also subtracted to obtain sets of genes found only within one of the lineages. Their distribution among chromosomes is indicated in the pie charts, and their functional distribution based on COG functional groupings is shown below. The COG categories are as described in the legend for Figure 2.4.

Discussion

An updated phylogeny of the *Burkholderia* genus

The use of the 16S rRNA gene for species discrimination has been one of the staple markers used by *Burkholderia* researchers. While other genes such as *recA* have also been used since they afford more phylogenetic discriminatory power, the multi-locus sequence typing method which probes portions of 6-7 loci has been the most powerful approach. However, different groups have selected different gene sets (e.g. (16, 40)) making comparisons between results difficult, and the choice of genes may confound results if they are frequently exchanged within the population (31), or are duplicated as has been discovered in one case recently (40). With the availability of many whole genomes, an even broader phylogenomic applications have been developed (9), whereby more than 30 loci are used, and have been shown to be universally distributed in bacteria, are present in one copy in most cases, and are recalcitrant to lateral gene transfer (LGT) (49). Using this phylogenomic approach, and later confirming the topology with a phylogeny derived from the core *Burkholderia* genome (all genes found in all *Burkholderia*), we have constructed an updated view of the *Burkholderia* genus.

The tree is consistent with respect to previous topologies for the environmental and rhizosphere isolates, but interestingly places the new *B. rhizoxinica* as a significantly distant outgroup to the other completed *Burkholderia* genomes. Whether this reflects its true phylogenetic position or simply an accelerated evolutionary rate partly determined by a genome reduction strategy (or a combination of both), remains a question that may best be answered with further genome sequencing of related isolates. Despite using multiple genes, either those shared throughout the tree of life, or those central to all *Burkholderia*,

the evolutionary signal within the *B. mallei* group appears to be too weak to robustly infer their evolutionary history. Perhaps using other genes present but specific to the Pseudomallei group or within the core of *B. mallei* genomes may best help reconstruct their phylogenetic relationships. However, this is provided that gene transfer and homologous recombination do not obscure such a reconstruction.

While the topology of the *Bc*c species has been described numerous times, we definitively place *B. glumae* is outside both the *Bc*c and Pseudomallei group. The *Burkholderia* gene core tree was highly congruent with the 31 gene tree, however two interesting differences were noted. The strongly supported placement of *B. ambifaria* as a sister taxa to *B. vietnamiensis* conflicts with the 31 gene tree, and the rearrangement of the, albeit, very recently diverged *B. pseudomallei* and *B. mallei* branches. Although effort was made to remove regions of uncertainty in the alignment, investigation of all the individual gene trees may help identify the source of this discrepancy. It is possible that a subset of the core is often exchanged among these strains and may mask the phylogenetic signal in the remaining core genes when concatenated together. Despite these small topological differences, given the bigger clade differences observed with many 16S rRNA phylogenies (with ours as well as among 16S rRNA tree topologies), and validation of the gross topology with the core 829 genes, we suggest using the topology presented here (Figure 2.1) to reflect evolutionary relationships between species.

Genome structure variation within the *Burkholderia* was reported over 15 years ago (8), and although it has been clear that many genomic islands have been acquired by any given species (7, 20, 30), whole genome alignments continue to reveal many interesting details regarding genome structure and similarity/differences between the various clades

within the *Burkholderia* genus. While the *Bc*c share three large replicons (all termed chromosomes), the other species within *Burkholderia* harbor two (though some strains do carry large plasmids, e.g. *B. phymatum* STM815 or *Burkholderia* sp. CCGE1002). In addition, while *B. mallei* is a recently derived clone of *B. pseudomallei*, it has undergone many genome rearrangements and large scale deletions most likely promoted by the >100 insertion sequences present in *B. mallei* (20). This is in contrast to the gross colinearity observed between members of the less closely related *Bc*c species or between *B. pseudomallei* and *B. thailandensis*.

While the rearrangements within *B. mallei* occur throughout and are attributed to intramolecular homologous recombination between insertion sequence elements, the other rearrangements observed between species almost always appear to occur around the origin or terminus of replication, as has been previously described (12). Further illustrated by the whole genome alignments is the more rapid sequence divergence of chromosome 2 compared with chromosome 1. Although closely related species maintain similar chromosome cohesion between chromosomes 1 and 2, more distant species preferentially maintain chromosome 1 cohesion, indicating a higher evolutionary rate in chromosome 2. The same holds true, or is even exacerbated, for the smaller replicons.

The *Burkholderia* genus pangenome

As may be expected given the plastic nature of the *Burkholderia* genome, the genus core consists of only 829 protein families, while the entire pangenome is more than eight times larger, at 56,777 protein families, including 43,247 "unique" protein families that are found in only a single genome. Despite such a large number, each genome carries on

average 26.7% unique genes, along with the core (13.3%) and the remaining 60% composed of the variable genome. The pangenome is thus essentially unlimited due in part to the role of LGT, bounded only by the variability in sequence space, or at least those sequences carried by accessible phage and other vehicles of LGT.

Interestingly, the core genome consists of genes that reside for the most part on chromosome 1, and these consist of translation functions and cell wall/cell division functions while the variable genes are found equally distributed among all replicons. Perhaps not surprisingly, the unique genome is derived mostly from either non-chromosome replicons, and from genomic islands of foreign origin that have taken residence within the main chromosome. It is possible, if not likely, that these unique segments of DNA are responsible for providing some strains or species with the ability to colonize new habitats, thus contributing to niche specialization. Among the many species-specific genes discovered here, many of them could be considered helpful or even essential in certain environmental habitats. In terms of broad functional category, the variable and unique genes are abundant in responding to environmental signals via signal transduction and transcriptional machinery and responding with carbohydrate and amino acid transport and metabolism functions.

The important role of species-specific genes

While it is clear that strain-specific genomic features can provide a particular strain with unique characteristics, including within the *Burkholderia* (6-7, 35), here we examined the unique gene content of several species and report relevant findings that may contribute to known and important phenotypes or to their lifestyles. Since the biothreat *Burkholderia*

and those involved in cystic fibrosis epidemics have been well described elsewhere, we focus here on previously unexplored features of the lesser known *Bc*c and non-*Bc*c, non-Pseudomallei group *Burkholderia*.

Some of these species and strains are well known for their xenobiotic degradation capabilities. While *B. vietnamiensis* G4 is known for its ability to degrade trichloroethylene and toluene (26, 29), *B. lata*, *B. phymatum* and *Burkholderia* sp. CCGE1002 also encode toluene tolerance or toluene-4-monooxygenase. In addition, *B. lata*, *Burkholderia* sp. CCGE1003, *B. phymatum*, *B. phytofirmans* and *B. rhizoxinica* also carry unique genes involved in organic solvent tolerance. Similarly *B. xenovorans* is known to carry many unique dioxygenases and other aromatic degradation genes (7), and a number of similar aromatic ring cleaving enzymes are also present in *B. phymatum*, *B. lata*, *Burkholderia* spp. CCGE1002 and CCGE1003, which may be useful during their close association with plants and plant-supplied organic compounds within the rhizosphere and plant detritus. In addition, *B. phytofirmans* has two unique nitrile and cyanate hydratases that may be involved in detoxification of xenobiotics and nitriles produced as defense chemicals by other microorganisms and plants, as well as in secondary metabolite biosynthetic pathways (36).

The *Burkholderia* are also renown for producing a number of important compounds, including many secondary metabolites via a variety of polyketide and non-ribosomal peptide synthesis pathways. While members of the *Bc*c, the Pseudomallei group and *B. glumae* all carry a number of unique polyketide and non-ribosomal peptide synthases, few if any are found within the "non-*Bc*c plant-associated" *Burkholderia*. Interestingly, *B. rhizoxinica* carries a large number of non-ribosomal peptide synthase modules, including

several on its plasmids that may in fact be responsible for the production of the rhizotoxin that affects plants (33). In addition, *B. rhizoxinica* carries a large family of unique hemolysins, which have often been implicated in bacterial pathogenesis. Further, *B. rhizoxinica* also carries a number of insecticidal proteins on one of its plasmids, which could potentially allow its fungal host to take advantage of the multiple toxins provided by its endosymbiont. Interestingly, both *B. phymatum* and *Burkholderia* sp. CCGE1002 encode a unique enterotoxin, however a role in plant symbiosis for enterotoxins has not yet been advanced, thus it is possible that this toxin may play an alternate role in perhaps a different host, or alternatively these species may also spend part of their lifecycle associated with other eukaryotic hosts.

Beyond toxin production, the *Burkholderia* are often considered plant growth promoting organisms. *B. phymatum*, the first non-alphaproteobacterial nodulator of plants (27) is known to fix nitrogen (14) and is the only genome in the studied group to carry a NifH nitrogenases as an isolated gene in addition to a *nifHDK* operon, while other *Burkholderia*, such as *B. xenovorans*, *B. vietnamiensis* and *Burkholderia* sp. CCGE1002 harbor only the operonic version, and several other *Burkholderia* spp. only carry nitrogenase cofactor biosynthesis proteins. A number of genes involved in production or tolerance to the phytohormone auxin are unique to either *B. glumae* or *B. phytofirmans*. In addition, *B. phytofirmans* possesses a unique urease, which could potentially play a role in antifungal growth inhibition of fungal phytopathogens (2). As with many soil organisms that may frequently encounter starvation conditions, *B. phytofirmans* and *B. phymatum* appear to store energy and carbon as biopolymer granules such as polyhydroxyalkanoate or polyhydroxybutyrate to be used in times of strife. Interestingly they each have one or

two unique phasin family proteins, known to be involved with biopolymer granules (21, 51-52).

Ankyrin repeat containing proteins are quite rare in bacteria , and while their function is unknown, they are believed to be involved in modulating infection of host cells (32, 46-47). Although a few ankyrin proteins can be found in the *Burkholderia* (*B. cenocepacia*, *B. glumae*, *B. lata* and *B. xenovorans*), there are a surprisingly large (>25) number of such proteins in *B. vietnamiensis*, some with only distant homologs in eukaryotes, and suggests a recent expansion of this gene family and possible new role in bacteria.

In addition to the above, each strain thus has its own set of unique proteins to interact with and respond to its environment, such as ABC and other transporters, efflux pumps, outer membrane proteins or other enzymes that decorate membrane lipid structures, pili, fimbriae, chemotaxis and flagellar proteins, siderophore receptors, heavy metal tolerance, heat and cold shock, as well as stress proteins, etc. Given these observations and the diversity of different environments in which *Burkholderia* thrive, we suggest that species- and possibly strain-specific genic content, and translated functional capabilities, may dictate the proficiency with which some bacteria colonize particular niches.

Differential genic content among lineages of the *Burkholderia*

Just as some species have specialized to, and thrive in particular environments, perhaps due in no small part to acquired functions that could allow expansion into a particular niche, lineages of several species within the *Burkholderia* may also be thought of

as successful clades that have since diversified and undergone speciation. We compared three main well-defined lineages, the nine *Bcc*, the nine Pseudomallei group, and the clade which harbors five soil rhizosphere dwellers and plant-associated *Burkholderia*. Although the pangenome for each of the lineages was roughly 20,000 genes in size, the core genomes were <3000 genes. When the three lineage core genomes were compared, roughly 500-600 genes remained lineage-specific for all three groups. These were further screened for the presence of even one ortholog in one of the two other lineages. With this depleted set of 180-350 genes, their functional assignments based on COG functional categories and their genomic locations were analyzed. A few functional categories appeared to be important for the lineages. For example, sugar and amino acid transport and metabolism genes were found among the *Bcc*-specific genes as well as among the rhizosphere lineage-specific genes. Also present in high abundance in the rhizosphere lineage core are signal transduction genes, as well as many genes with unknown function. This was also true for the Pseudomallei group, along with genes for energy production and conversion. Given the differences observed in rates of evolution between the chromosomes, it was interesting to note that the distribution of these lineage-specific genes in the genome differed substantially between lineages, with the Bcc harboring most of their lineage-specific core on chromosome 2, the rhizosphere lineage primarily on chromosome 1, and the Pseudomallei lineage was relatively even between its two chromosomes.

Summary

The *Burkholderia* are a highly diverse and versatile group of bacteria, matched by their genomic complexity and their flexible evolutionary strategies. The observed genomic

rearrangements among species and the differences in rates of sequence divergence (evolution) among replicons (chromosomes and plasmids) suggest a number of different modes of evolution at work within this species. The division or partitioning of core, variable and unique genes among their differentially evolving replicons allows these organisms a measure of flexibility in uptake and "sampling" of foreign material (through novel acquisition, deletion, or mutation), and may permit rapid adaptation to newly available niches and offer a competitive advantage. It is likely that a complex combination of local environment, lifestyle, and other selective pressures, along with genomic content, and historical evolutionary strategy may dictate the evolutionary trajectory of these diverse bacteria.

References

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-10.

2. **Becker-Ritt, A. B., A. H. Martinelli, S. Mitidieri, V. Feder, G. E. Wassermann, L. Santi, M. H. Vainstein, J. T. Oliveira, L. M. Fiuza, G. Pasquali, and C. R. Carlini.** 2007. Antifungal activity of plant and bacterial ureases. Toxicon **50:**971-83.

3. **Burkholder, W. H.** 1942. Three bacterial plant pathogens. Phytomonas caryophylli sp.n., Phytomonas alliicola sp.n. and Phytomonas manihotis Phytopathology **32:**141-149.

4. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis Comparison Tool. Bioinformatics **21:**3422-3.

5. **Castillo, D. M., and T. E. Pawlowska.** 2010. Molecular evolution in bacterial endosymbionts of fungi. Mol Biol Evol **27:**622-36.

6. **Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia.** 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A **101:**13826-31.

7. **Chain, P. S., V. J. Denef, K. T. Konstantinidis, L. M. Vergez, L. Agullo, V. L. Reyes, L. Hauser, M. Cordova, L. Gomez, M. Gonzalez, M. Land, V. Lao, F. Larimer, J. J. LiPuma, E. Mahenthiralingam, S. A. Malfatti, C. J. Marx, J. J. Parnell, A. Ramette, P. Richardson, M. Seeger, D. Smith, T. Spilker, W. J. Sul, T. V. Tsoi, L. E. Ulrich, I. B. Zhulin, and J. M. Tiedje.** 2006. Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A **103:**15280-7.

8. **Cheng, H. P., and T. G. Lessie.** 1994. Multiple replicons constituting the genome of Pseudomonas cepacia 17616. J Bacteriol **176:**4034-42.

9. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283-7.

10. **Coenye, T., P. Vandamme, J. R. Govan, and J. J. LiPuma.** 2001. Taxonomy and identification of the Burkholderia cepacia complex. J Clin Microbiol **39:**3427-36.

11. **Cooper, V. S., S. H. Vohr, S. C. Wrocklage, and P. J. Hatcher.** 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput Biol **6:**e1000732.

12. **Deng, W., V. Burland, G. Plunkett, 3rd, A. Boutin, G. F. Mayhew, P. Liss, N. T. Perna, D. J. Rose, B. Mau, S. Zhou, D. C. Schwartz, J. D. Fetherston, L. E. Lindler, R. R. Brubaker, G. V. Plano, S. C. Straley, K. A. McDonough, M. L. Nilles, J. S. Matson, F. R. Blattner, and R. D. Perry.** 2002. Genome sequence of Yersinia pestis KIM. J Bacteriol **184:**4601-11.

13. **Edgar, R. C.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:**113.

14. **Elliott, G. N., W. M. Chen, J. H. Chou, H. C. Wang, S. Y. Sheu, L. Perin, V. M. Reis, L. Moulin, M. F. Simon, C. Bontemps, J. M. Sutherland, R. Bessi, S. M. de Faria, M. J. Trinick, A. R. Prescott, J. I. Sprent, and E. K. James.** 2007. Burkholderia phymatum is a highly effective nitrogen-fixing symbiont of Mimosa spp. and fixes nitrogen ex planta. New Phytol **173:**168-80.

15. **Fulton, D. L., Y. Y. Li, M. R. Laird, B. G. Horsman, F. M. Roche, and F. S. Brinkman.** 2006. Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics **7:**270.

16. **Gee, J. E., M. B. Glass, G. Lackner, L. O. Helsel, M. Daneshvar, D. G. Hollis, J. Jordan, R. Morey, A. Steigerwalt, and C. Hertweck.** 2011. Characterization of Burkholderia rhizoxinica and B. endofungorum Isolated from Clinical Specimens. PLoS One **6:**e15731.

17. **Glass, M. B., J. E. Gee, A. G. Steigerwalt, D. Cavuoti, T. Barton, R. D. Hardy, D. Godoy, B. G. Spratt, T. A. Clark, and P. P. Wilkins.** 2006. Pneumonia and septicemia caused by Burkholderia thailandensis in the United States. J Clin Microbiol **44:**4601-4.

18. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696-704.

19. **Ho, C. C., C. C. Lau, P. Martelli, S. Y. Chan, C. W. Tse, A. K. Wu, K. Y. Yuen, S. K. Lau, and P. C. Woo.** 2010. A novel pan-genomic analysis approach in target selection for multiplex PCR identification and detection of Burkholderia pseudomallei, Burkholderia thailandensis and Burkholderia cepacia complex species: a proof-of-concept study. J Clin Microbiol.

20. **Holden, M. T., R. W. Titball, S. J. Peacock, A. M. Cerdeno-Tarraga, T. Atkins, L. C. Crossman, T. Pitt, C. Churcher, K. Mungall, S. D. Bentley, M. Sebaihia, N. R. Thomson, N. Bason, I. R. Beacham, K. Brooks, K. A. Brown, N. F. Brown, G. L. Challis, I. Cherevach, T. Chillingworth, A. Cronin, B. Crossett, P. Davis, D. DeShazer, T. Feltwell, A. Fraser, Z. Hance, H. Hauser, S. Holroyd, K. Jagels, K. E. Keith, M. Maddison, S. Moule, C. Price, M. A. Quail, E. Rabbinowitsch, K. Rutherford, M. Sanders, M. Simmonds, S. Songsivilai, K. Stevens, S. Tumapa, M. Vesaratchavest, S. Whitehead, C. Yeats, B. G. Barrell, P. C. Oyston, and J. Parkhill.** 2004. Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proc Natl Acad Sci U S A **101:**14240-5.

21. **Jurasek, L., and R. H. Marchessault.** 2004. Polyhydroxyalkanoate (PHA) granule formation in Ralstonia eutropha cells: a computer simulation. Appl Microbiol Biotechnol **64:**611-7.

22. **Katoh, K., and H. Toh.** 2010. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics **26:**1899-900.

23. **Lackner, G., N. Moebius, L. Partida-Martinez, and C. Hertweck.** 2011. Complete genome sequence of Burkholderia rhizoxinica, an Endosymbiont of Rhizopus microsporus. J Bacteriol **193:**783-4.

24. **Losada, L., C. M. Ronning, D. DeShazer, D. Woods, N. Fedorova, H. S. Kim, S. A. Shabalina, T. R. Pearson, L. Brinkac, P. Tan, T. Nandi, J. Crabtree, J. Badger, S. Beckstrom-Sternberg, M. Saqib, S. E. Schutzer, P. Keim, and W. C. Nierman.** 2010. Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. Genome Biol Evol **2:**102-16.

25. **Mahenthiralingam, E., T. A. Urban, and J. B. Goldberg.** 2005. The multifarious, multireplicon Burkholderia cepacia complex. Nat Rev Microbiol **3:**144-56.

26. **Mars, A. E., J. Houwing, J. Dolfing, and D. B. Janssen.** 1996. Degradation of Toluene and Trichloroethylene by Burkholderia cepacia G4 in Growth-Limited Fed-Batch Culture. Appl Environ Microbiol **62:**886-91.

27. **Moulin, L., A. Munive, B. Dreyfus, and C. Boivin-Masson.** 2001. Nodulation of legumes by members of the beta-subclass of Proteobacteria. Nature **411:**948-50.

28. **Nandi, T., C. Ong, A. P. Singh, J. Boddey, T. Atkins, M. Sarkar-Tyson, A. E. Essex-Lopresti, H. H. Chua, T. Pearson, J. F. Kreisberg, C. Nilsson, P. Ariyaratne, C. Ronning, L. Losada, Y. Ruan, W. K. Sung, D. Woods, R. W. Titball, I. Beacham, I. Peak, P. Keim, W. C. Nierman, and P. Tan.** 2010. A genomic survey of positive

selection in Burkholderia pseudomallei provides insights into the evolution of accidental virulence. PLoS Pathog **6:**e1000845.

29. **Nelson, M. J., S. O. Montgomery, W. R. Mahaffey, and P. H. Pritchard.** 1987. Biodegradation of trichloroethylene and involvement of an aromatic biodegradative pathway. Appl Environ Microbiol **53:**949-54.

30. **Nierman, W. C., D. DeShazer, H. S. Kim, H. Tettelin, K. E. Nelson, T. Feldblyum, R. L. Ulrich, C. M. Ronning, L. M. Brinkac, S. C. Daugherty, T. D. Davidsen, R. T. Deboy, G. Dimitrov, R. J. Dodson, A. S. Durkin, M. L. Gwinn, D. H. Haft, H. Khouri, J. F. Kolonay, R. Madupu, Y. Mohammoud, W. C. Nelson, D. Radune, C. M. Romero, S. Sarria, J. Selengut, C. Shamblin, S. A. Sullivan, O. White, Y. Yu, N. Zafar, L. Zhou, and C. M. Fraser.** 2004. Structural flexibility in the Burkholderia mallei genome. Proc Natl Acad Sci U S A **101:**14246-51.

31. **Ochman, H., J. G. Lawrence, and E. A. Groisman.** 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405:**299-304.

32. **Pan, X., A. Luhrmann, A. Satoh, M. A. Laskowski-Arce, and C. R. Roy.** 2008. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. Science **320:**1651-4.

33. **Partida-Martinez, L. P., and C. Hertweck.** 2005. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. Nature **437:**884-8.

34. **Perin, L., L. Martinez-Aguilar, G. Paredes-Valdez, J. I. Baldani, P. Estrada-de Los Santos, V. M. Reis, and J. Caballero-Mellado.** 2006. Burkholderia silvatlantica sp. nov., a diazotrophic bacterium associated with sugar cane and maize. Int J Syst Evol Microbiol **56:**1931-7.

35. **Perna, N. T., G. Plunkett, 3rd, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner.** 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature **409:**529-33.

36. **Podar, M., J. R. Eads, and T. H. Richardson.** 2005. Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. BMC Evol Biol **5:**42.

37. **Price, M. N., P. S. Dehal, and A. P. Arkin.** 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One **5:**e9490.

38. **Sessitsch, A., T. Coenye, A. V. Sturz, P. Vandamme, E. A. Barka, J. F. Salles, J. D. Van Elsas, D. Faure, B. Reiter, B. R. Glick, G. Wang-Pruski, and J. Nowak.** 2005. Burkholderia phytofirmans sp. nov., a novel plant-associated bacterium with plant-beneficial properties. Int J Syst Evol Microbiol **55:**1187-92.

39. **Sim, S. H., Y. Yu, C. H. Lin, R. K. Karuturi, V. Wuthiekanun, A. Tuanyok, H. H. Chua, C. Ong, S. S. Paramalingam, G. Tan, L. Tang, G. Lau, E. E. Ooi, D. Woods, E. Feil, S. J. Peacock, and P. Tan.** 2008. The core and accessory genomes of Burkholderia pseudomallei: implications for human melioidosis. PLoS Pathog **4:**e1000178.

40. **Spilker, T., A. Baldwin, A. Bumford, C. G. Dowson, E. Mahenthiralingam, and J. J. LiPuma.** 2009. Expanded multilocus sequence typing for burkholderia species. J Clin Microbiol **47:**2607-10.

41. **Sun, L., R. Z. Jiang, S. Steinbach, A. Holmes, C. Campanelli, J. Forstner, U. Sajjan, Y. Tan, M. Riley, and R. Goldstein.** 1995. The emergence of a highly transmissible lineage of cbl+ Pseudomonas (Burkholderia) cepacia causing CF centre epidemics in North America and Britain. Nat Med **1:**661-6.

42. **Talavera, G., and J. Castresana.** 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol **56:**564-77.

43. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

44. **Thompson, J. D., T. J. Gibson, and D. G. Higgins.** 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics **Chapter 2:**Unit 2 3.

45. **Ussery, D. W., K. Kiil, K. Lagesen, T. Sicheritz-Ponten, J. Bohlin, and T. M. Wassenaar.** 2009. The genus burkholderia: analysis of 56 genomic sequences. Genome Dyn **6:**140-57.

46. **Voronin, D. A., and E. V. Kiseleva.** 2007. [Functional role of proteins containing ankyrin repeats]. Tsitologiia **49:**989-99.

47. **Walker, T., L. Klasson, M. Sebaihia, M. J. Sanders, N. R. Thomson, J. Parkhill, and S. P. Sinkins.** 2007. Ankyrin repeat domain-encoding genes in the wPip strain of Wolbachia from the Culex pipiens group. BMC Biol **5:**39.

48. **Weinberg, J. B., B. D. Alexander, J. M. Majure, L. W. Williams, J. Y. Kim, P. Vandamme, and J. J. LiPuma.** 2007. Burkholderia glumae infection in an infant with chronic granulomatous disease. J Clin Microbiol **45:**662-5.

49. **Wu, M., and J. A. Eisen.** 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol **9:**R151.

50. **Yabuuchi, E., Y. Kosako, H. Oyaizu, I. Yano, H. Hotta, Y. Hashimoto, T. Ezaki, and M. Arakawa.** 1992. Proposal of Burkholderia gen. nov. and transfer of seven species of the genus Pseudomonas homology group II to the new genus, with the type species Burkholderia cepacia (Palleroni and Holmes 1981) comb. nov. Microbiol Immunol **36:**1251-75.

51. **York, G. M., B. H. Junker, J. A. Stubbe, and A. J. Sinskey.** 2001. Accumulation of the PhaP phasin of Ralstonia eutropha is dependent on production of polyhydroxybutyrate in cells. J Bacteriol **183:**4217-26.

52. **York, G. M., J. Stubbe, and A. J. Sinskey.** 2001. New insight into the role of the PhaP phasin of Ralstonia eutropha in promoting synthesis of polyhydroxybutyrate. J Bacteriol **183:**2394-7.

Chapter 3.
Species genomes and genome representatives: Insights from comparative Pangenomics of *Burkholderia* species.

Introduction

With the introduction of whole genome sequencing in 1995 (23) and the subsequent and continually growing flood of genome projects, the issue of "How representative is a single genome?" has arisen and become a topic of great interest. The true question however, is "representative of what?"! A single genome certainly cannot be considered representative for an entire phylum, nor for that matter a genus, but does a single genome represent a species? Enter the perennial debate on a bacterial species concept (19-20, 34, 74).

Although bacteriologists have not adopted a universal bacterial species concept (19, 76), the widely accepted current bacterial species definition specifies strains as belonging to the same species if they share distinctive phenotypic properties from near relatives as well as have greater than 70% DNA-DNA hybridization (86). Although this definition is somewhat arbitrary (and artificial), it has helped standardize bacterial and archaeal taxonomy (72, 75). A number of exceptions arise given any definition of bacterial species, and so the DNA-DNA relatedness criterion has been called into question for being too stringent in defining some species (84), while being too broad in defining others (75). In the age of sequencing, more natural and expedient approaches to define species has been sought (reviewed in (26)).

The comparison of DNA sequences, particularly the ubiquitously found 16S ribosomal RNA (rRNA) molecule, is one approach proposed to enhance the definition of bacterial species (73). Shown to be highly correlated with DNA-DNA reassociation studies, a rRNA identity of less than 97% is generally considered to reflect different species, and this cutoff is often used in defining operational taxonomic units in environmental microbial

diversity studies (1, 15, 41-42). While this description has undergone revision (71), it has been previously shown that the 16S rRNA may not provide sufficient discriminatory power, since organisms sharing nearly identical 16S rRNA genes may differ greatly in genome content or the remainder of their shared genomic sequences and should not necessarily be classified as the same species (43, 67).

Thus, while still invaluable as a measure of diversity in metagenomic studies, a number of more suitable methods have since been used for resolving closely related species and defining subpopulations within species. These include sequence analysis of more rapidly evolving protein-coding sequences such as the *recA* gene, and rapid DNA amplification and/or restriction methods such as restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), randomly amplified polymorphic DNA (RAPD), repetitive extragenic palindromic sequence (REP)-PCR, to name a few. In an extension of the single gene analysis, multi-locus sequence typing (52) and multi-locus sequence analysis (26), which target a number of housekeeping genes have also been used to analyze both intra- and inter-species relationships, such as in the *Burkholderia* (3, 27). These methods have been argued as good guides for species delineations, even in the face of highly recombinogenic clades (32-33). The application of a combined strategy using RecA and MLST for *Burkholderia* was shown to provide even greater resolving power than either technique alone (9).

The idea of using conserved housekeeping genes for understanding the evolutionary histories and relatedness of species is not a new one (85), and now with the large number of genomes available, new ideas on conserved gene sets to be used for such analyses have emerged. This brings us back to the question of whether a single genome can represent a

species. Although genome size and structure were known to vary among phylogenetically closely related isolates, including *Escherichia coli* (5) and *Burkholderia* spp. (51), the extent of genomic plasticity was first examined in detail with three genomes of *E. coli* (87). Thus it is clear that one genome is insufficient to represent a "species", despite only contributing minimally to the concept of a bacterial species. This mosaicism is due in large part to the acquisition and maintenance of novel genomic islands, introduced via lateral gene transfer (LGT). LGT has been claimed to be rampant across all classes of genes, causes complications in reconstructing phylogenetic histories due to homologous recombination, and is a prime culprit in obfuscating a bacterial species concept (e.g. (4, 18-20, 28, 60, 88).

The notion of a conserved core of genes resistant to LGT has been proposed (14, 40), and its use in defining a "species genome" was presented by Lan and Reeves over a decade ago (47). More recently, the term "pangenome" was coined by Tettelin et al. (2005), and defined as the total number of genes found in at least one genome within a species (24, 79). This includes the core genome, shared by all (or most) members of a species, the variable (also termed auxiliary, adaptive or character) genome, shared by some but not all members of a species, and the unique (or accessory) genome, present in only one (or a small minority) of the species. The authors attempted to track the number of genomes necessary to capture the entire pangenome and concluded that two types of pangenomes existed, an open pangenome with an unlimited number of genes, and a closed pangenome, where a finite number of individual sequences would be required to describe the species pangenome. However, it seems clear that the distinction has more to do with the definition of species than the concept of a species.

The pangenome concept may still enlighten our concept of species. For example, the mere presence or absence of pangenome (core and variable) genes distributed within species has been shown to provide higher discriminating resolution than MLST (31). A similar, yet more elegant approach was developed to distinguish closely related strains via the average nucleotide identity of their shared genes, defined using an empirical set of cutoff values (43). However, it is clear that a comprehensive interpretation of cohesive biological groups ("species") will require analysis of the entire pangenome, including non-core genes (53). Indeed, when this pangenome concept was applied to the entire bacterial domain, although most gene families were part of the unique or accessory gene pool (present in a small minority of genomes), the majority of genes within any given bacterial genome was comprised of the variable genome (not the accessory genome), present in a number of different bacteria (48).

Since the analysis of several strains of *Streptococcus agalactiae* (79), a rising number of pangenome studies for other pathogenic and model species are being reported, including for *Streptococcus pneumonia* (35), *E. coli* (63), *Vibrio cholera* (80). There has also been a recent flurry of reports on genus pangenomes that range in scope from simply describing methods and size of pangenome components of *Burkholderia* spp. and *Brucella* spp. (6, 83) to a comparative description of differential gene composition and genome evolution among *Listeria* spp. and *Bifidobacterium* spp. (7, 16). New tools and methods are also being developed for this relatively new subfield, from methods to predict orthologs (82), to software to help process pangenome queries of highly related strains (46).

Here, we present a novel pangenomic framework for looking at *Burkholderia* species and discuss how this can be extended to interpret the diversity and evolutionary forces

driving the formation of other cohesive groups of bacteria. The *Burkholderia* spp. have been a key focus of many studies involving taxonomic classification, in part due to the historic use of some of its species as bioweapons, and the prevalence of other species in lethal necrotizing pneumonia among cystic fibrosis patients. This interesting and phenotypically diverse group of species is also found ubiquitously in the environment, and many species form favorable symbioses with eukaryotic hosts. The genomic homogeneity within certain clades has often amalgamated many species together, and combined with the genomic heterogeneity between other clades, this provides a wonderful opportunity to compare the pangenomes of genomically similar as well as genomically distinct species within a single genus. A number of studies have already been performed that have discriminated between *Burkholderia* isolates using 16S rRNA comparisons, and perhaps more successfully using either faster evolving single marker genes such as RecA (62), or by looking at genome-wide patterns such as MLST (3), average nucleotide identity (42), gene presence or absence (83). This is the first comparative pangenomic study of genome-differentiating factors that may contribute to the cohesiveness and distinctiveness of 'species'. Trends in species pangenomes within the *Burkholderia* genus are presented along with possible species rules or trends as observed in this genus.

Materials and Methods

Burkholderia species and strains

For this study, we analyzed all genomes for which there were 2 or more completed genomes within a single designated species of *Burkholderia* (Table 3.1) thus restricting our study to *B. cenocepacia* (4 strains), *B. pseudomallei* (4), *B. mallei* (4) and *B. ambifaria* (2). While there are several *Burkholderia* species for which there exist two or more genome projects with draft data available (*B. multivorans*, *B. oklahomensis*, *B. pseudomallei*, *B. mallei*, *B. ambifaria*, *B. thailandensis*), the projects are composed of many contiguous sequences (or contigs), a sign of poor assembly of raw data that number up to >2300 contigs, and reflects poor genome coverage and/or poor quality. Since these could result in a number of annotation anomalies due to sequencing errors, they were not included.

Phylogenetic tree construction

Thirty-one conserved 'housekeeping' proteins were retrieved from each genome using preconstructed hidden Markov models in AMPHORA (89). MUSCLE (v3.8.31) was used to construct a multiple sequence alignment of the 31 concatenated protein sequences (21). Poorly conserved positions with obscure ancestry were excluded from phylogenetic analyses using Gblocks ver.0.91b (77) to select conserved blocks (Settings: 14 for Minimum Number Of Sequences For A Conserved Position and For A Flank Position; 15 for Maximum Number Of Contiguous Nonconserved Positions; 3 for Minimum Length Of A Block; and 50% for Allowed Gap Positions). A maximum likelihood phylogenetic tree was constructed using PHYML v3.0.1 with the Jones-Taylor-Thorton (JTT) model of amino acid substitution, and 100 bootstrapped replicates (29).

Table 3.1. Names and descriptions of strains selected for sequencing.*

Species	Size	G+C	# of Chrs (Pds)	Isolation	Genome Center (Release Date)
***B. ambifaria* AMMD**	7.52	66.8	3 (1)	Soil	JGI (2006)
***B. ambifaria* MC40-6**	7.64	66.4	3 (1)	Soil	JGI (2008)
***B. cenocepacia* AU 1054**	7.28	66.9	3	Soil	JGI (2006)
***B. cenocepacia* HI2424**	7.72	66.8	3 (1)	Clinical	JGI (2006)
***B. cenocepacia* J2315**	8.04	66.9	3 (1)	Clinical	WTSI (2008)
***B. cenocepacia* MC0-3**	7.96	66.6	3	Soil	JGI (2008)
***B. mallei* ATCC 23344**	5.83	68.5	2	Clinical	TIGR (2004)
***B. mallei* NCTC 10229**	5.76	68.5	2	Clinical	TIGR (2007)
***B. mallei* NCTC 10247**	5.85	68.5	2	Clinical	TIGR (2007)
B. mallei* SAVP1	5.23	68.4	2	Clinical	TIGR (2007)
***B. pseudomallei* 1106a**	7.10	68.3	2	Clinical	TIGR (2007)
***B. pseudomallei* 1710b**	7.30	68.0	2	Clinical	TIGR (2005)
***B. pseudomallei* 668**	7.03	68.3	2	Clinical	TIGR (2007)
***B. pseudomallei* K96243**	7.25	68.1	2	Clinical	WTSI (2004)

*Chrs, chromosomes; Pds, plasmids; JGI, Joint Genome Institute; WTSI, Welcome Trust Sanger Institute; TIGR, The Institute for Genomic Research; SAVP1 is classified as avirulent

Whole genome alignments and visualization

The fasta sequences of all replicons within all the genomes within this study were concatenated in order of chromosome, followed by a plasmid, if present. BlastN (2) was used to perform local alignments between selected pairs of genomes, and later filtered by E value, bit score and size of hit, with precise parameters that varied based on the genetic relatedness of each pairwise comparison. The visualization program ACT (8) was used to display each pairwise alignments.

Detailed comparative genomics within species

Whole genome sequences were aligned pairwise to each other using the NUCmer tool to find Maximal Unique Matches (MUMs), available from the MUMmer package ((45); http://mummer.sourceforge.net/). Multiple alignments per sequence were allowed to account for repeat regions, including insertion elements and other highly conserved sequences. Alignment files were then parsed with the show-snps and show-coords MUMmer tools to generate flat text files containing either Single Nucleotide Polymorphism (SNP) information, or the location, length and identity of any alignments. Coordinates files were parsed to find 'valid' alignments, which were calculated requiring a default cutoff of 94% identity over 500 bp. Gaps between valid alignments were assumed to exist from the end of any alignment to the beginning of the next. SNP calculations are performed internally to the MUMmer package, but this is a simple process for pair-wise alignments. Small insertions and deletions (indels) are calculated in areas where small stretches of reference does not match to the query. To identify all genes that are altered, missing or unique to any of the genomes, both SNP+indel and gap information were extracted and saved for additional analysis.

Gap or SNP coordinates were compared to the appropriate Genbank reference files, and genes contained within gaps, or containing one or more SNPs or indels were extracted. SNPs and indels in non-coding regions were not analyzed. The position of genes, location of SNPs or indels, if applicable, and the name of the gene were recorded. Additionally, in the case of SNPs, the nature of the base change (synonymous or non-synonymous) was also noted. This data was further collated into tables by reference species. The online Venn

diagram program Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html) was used to construct Venn diagrams illustrating overlaps among strains.

The COG functional categories of all gap, SNP, and indel containing genes were calculated from Genbank files for each organism or obtained from the COG database (78). Genes containing non-synonymous and indels were combined and these lists were used to determine COG functional breakdowns of genes affected by various mutations, insertions or deletions among members of the same species.

Evolutionary rate analysis

To estimate whether there were indications of purifying or positive selection, results from NUCmer (see above) genome alignments were parsed to identify those codons where SNPs occurred between any two strains of the same species. These codons within all the strains were then recorded, and the set of codons for each chromosome of each strain was then used as a concatenated fasta file of aligned codons, compared with those from other strains. The Synonymous Non-synonymous Analysis Program (44) was used to estimate synonymous (dS) and non-synonymous (dN) substitutions per site, with the Nei and Gojobori method (57) accounting for variances, covariances, and the possibility of multiple substitutions (59).

Core genome and pangenome analyses

Orthologs were identified by the program Ortholuge (25). Briefly, the reciprocal-best-blast hits were determined for all genes in each genome using a many-to-many blastP search (2) with an expectation value (E-value) cutoff of 1e-4. Each candidate was then

aligned with the others using ClustalW (81) and their phylogenetic distance ratios calculated with fprotdist, PHYLIP's protdist software compiled with EMBOSS as an EMBASSY package. Orthologs between genomes were then identified using ratio cutoff values of R1 ≤ 0.55 and R2 ≤ 0.70, as recommended. Orthologs not shared with the outgroup were detected using a reciprocal best blast solution. BLAST hits for each gene were collected if the top BLAST hit (cutoff of 1e-20) and at least 70% of the sequence in the alignment and with 60% or greater percent identity. If the top BLAST hit for a gene in genome A is the top BLAST hit of a gene in genome B, then a paired match is formed and these genes are designated orthologs. COG classifications (78) were assigned to each gene, if available, by parsing the PID from the *.ptt file that resides in the genome project FTP site at NCBI. Lookup tables were created to link the NC numbered files with the GI, COG, chromosome number, and sequence information (obtained by the *.faa files).

Results and Discussion

Whole genome alignments reveal flexible nature of species genomes

All *Burkholderia* species for which more than one genome was available were selected for pangenome analyses: two *B. ambifaria* strains, and four strains each of *B. cenocepacia*, *B. pseudomallei* and *B. mallei*. These belong to two distinct clades of *Burkholderia*, one with the *Bcc* members *B. ambifaria* and *B. cenocepacia*, and one with the biothreat agents *B. pseudomallei* and its recently evolved clone, *B. mallei* (Figure 3.1).

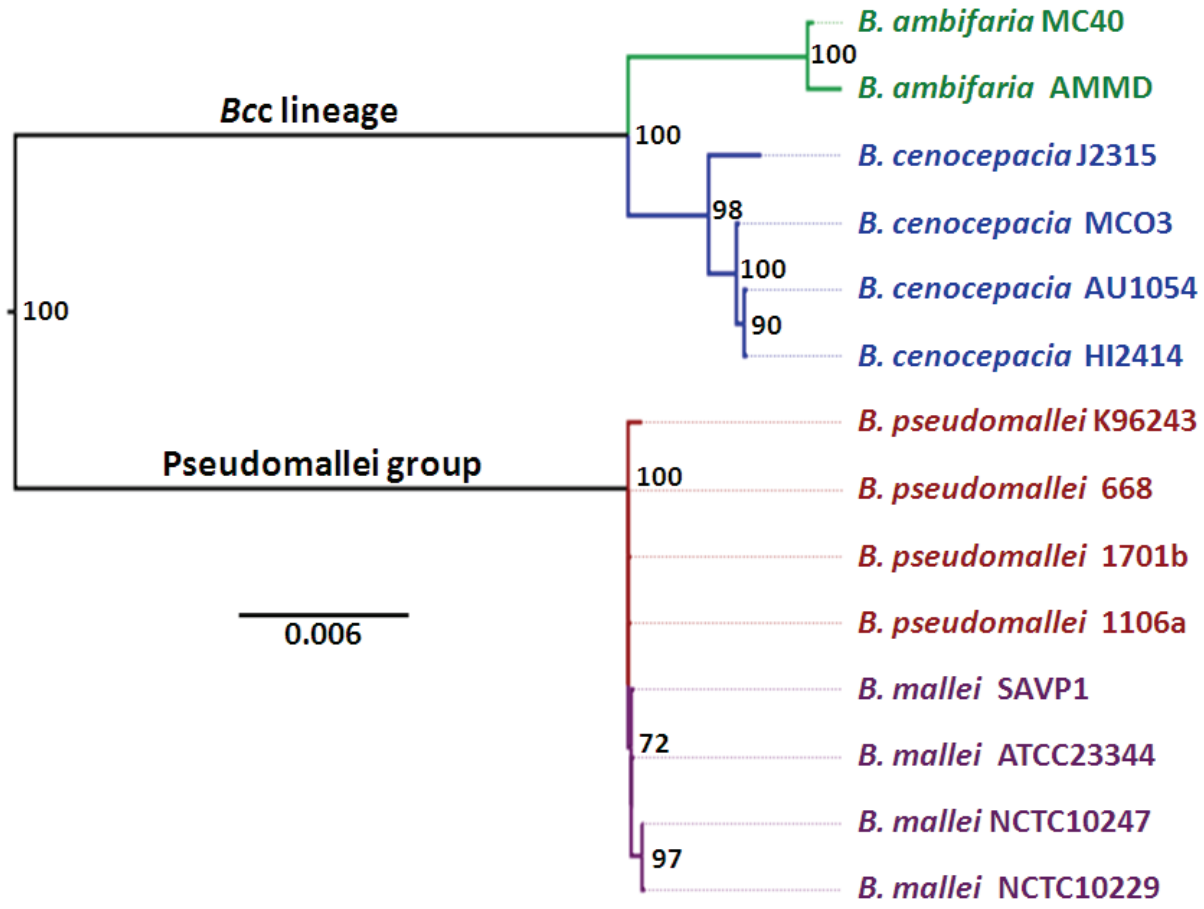


Figure 3.1. Phylogeny of strains within four *Burkholderia* species. A maximum likelihood phylogenetic tree derived from 31 housekeeping genes obtained from two *B.*

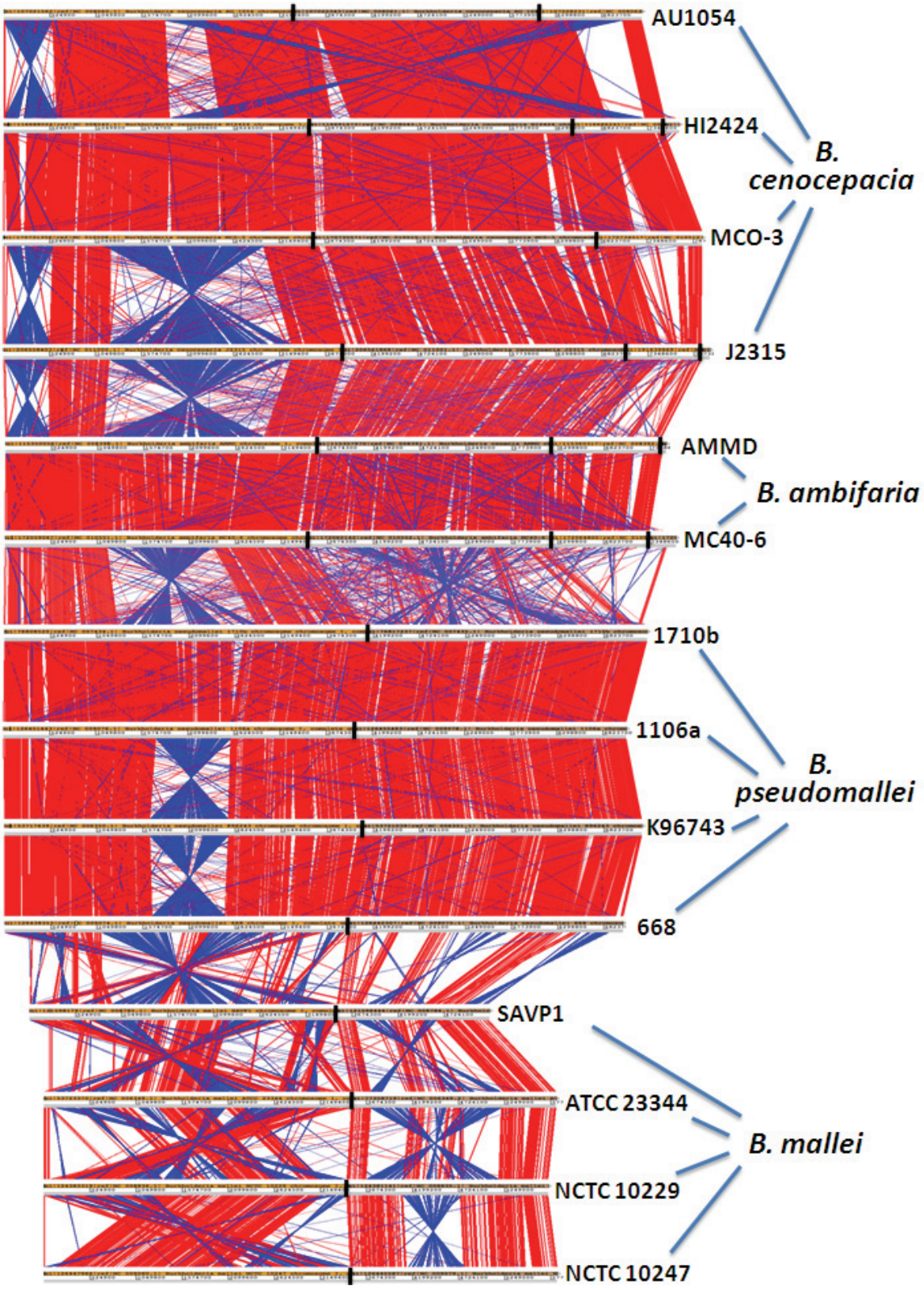
ambifaria (green), four *B. cenocepacia* (blue), four *B. pseudomallei* (maroon), and four *B. mallei* (purple) strains. Bootstrap values above 50 are shown. The Pseudomallei group remains poorly resolved. Scale bar represents substitutions per amino acid.

Gross structural variations were surveyed using whole genome alignments between all strains interrogated in this study (Figure 3.2). Interestingly, there are three chromosomes that are conserved among members of the *Bc*c instead of two for other members of *Burkholderia* (more in Chapter 2). Despite a well conserved backbone within each species, a number of genomic islands (and entire plasmids) present in some but not other strains of the same species are apparent, and are interpreted as either insertions of novel genetic material, deletions, or regions of sufficient divergence that nucleotide similarity is too low to be observed. There are also a number of strain-specific rearrangements that have accrued within a species, with the highly malleable genome of *B. mallei* representing an extreme case, where rearrangements are known to be frequent due to the abundance of IS elements in their genomes (37). Given that most intra-genome rearrangements occur about the origin/terminus of replication, one interesting observation is the inter-molecular translocation between chromosomes 1 and 3 of *B. cenocepacia* AU1054 (30). Although sequence similarity of this DNA region has been maintained, such changes in genomic context may affect important phenotypes such as growth rate or pathogenicity due to gene dosage or expression level effects (54, 65-66). Lastly, though there is a high degree of whole genome synteny and sequence similarity between the highly related *Bc*c species as well as between *B. pseudomallei* and its recently evolved clonal lineage *B. mallei*, there is moderate sequence divergence resulting in a

breakdown of sequence similarity in chromosome 2, as shown between the *Bcc* and the Pseudomallei group. This faster evolutionary rate has previously been observed in secondary chromosomes of *Burkholderia* and is reported elsewhere (10, 12). It has also been proposed that this more rapid sequence evolution on secondary chromosomes is related to a weaker codon usage bias and possibly reduced gene dosage and expression (12). A closer analysis of the genes found in the translocation between chromosome 1 and 3 in AU1054 could help test this hypothesis.

Figure 3.2 (*next page*). Comparing the structural pangenome of four *Burkholderia* species. The genomes four *B. cenocepacia*, two *B. ambifaria*, four *B. pseudomallei*, and four *B. mallei* genomes were concatenated, such that all chromosomes and plasmids (orange and brown horizontal boxes separated by black lines) were placed end to end. The strain names are indicated to the right of the concatenated sequence. Alignments between genomes are represented in the form of syntenic collinear blocks that are either parallel (red) or inverted (blue). (*Figure 3.2 on next page*)

Figure 3.2 cont'd.



Detailed comparative genomics reveals a range of differences within defined species

Due to the overall conserved nature of genomes within species, we sought to perform an even more refined nucleotide-level analysis that included identification of both large differences (e.g. genomic islands and deletions of one or more genes), as well as small differences such as SNPs or small (<20bp) insertions or deletions (indels). Through detailed nucleotide alignments, all highly similar DNA segments were compared and large regions of difference were examined further. Relating to the pangenome, which is defined as the total set of genes found within the genomes of the same species (24, 79), we identified all unique sequences present in one strain and not found in another strain of the same species (Table 3.2). While trends in the pangenomes can be developed from this analysis in terms of number of unique regions, amount of unique DNA, or number of unique genes each strain harbors, both the small number of genomes (n=2-4) and the non-randomness of selected strains for sequencing, may lead to inaccurate interpretations of the results since the strains may not be reflective of the species diversity (ie. are not randomly selected genomes). Nevertheless, many insights into the nature of *Burkholderia* species may still be gleaned, thus both a detailed analysis as well as a standard gene/protein-based pangenomic analysis were performed. Both methods gave similar statistics in terms of both shared and strain-specific gene numbers.

The most genetically diverse species among the four tested was *B. cenocepacia*, despite having two highly similar strains of the same clonal (PHDC) lineage (AU1054 has less than 4kb of unique sequence compared with HI2424). Both other sequenced *B. cenocepacia* strains, J2315 and MCO-3 made up for this with a minimum of almost 1 to 1.7 Mb of unique sequence when compared with any of the other strains. Similarly, each of the

two *B. ambifaria* strains harbored almost 200 regions with DNA totaling over 1.1Mb not present in the other strain. This contrasts with both *B. mallei* and *B. pseudomallei* strains who each have fewer unique regions with respect to other strains in the species and correspondingly, a smaller total amount of unique DNA and unique genes.

Table 3.2. Strain-specific regions in pairwise genome comparisons among species.*

B. cenocepacia **Unique to \ vs**	**vs AU1054**	**vs HI2424**	**vs J2315**	**vs MC0-3**
AU1054	-	3 (3,395)	347 (1,208,560)	154 (498,089)
HI2424	10 (426,268)	-	357 (1,422,845)	160 (713,727)
J2315	383 (1,842,710)	402 (1,727,228)	-	399 (1,675,856)
MC0-3	161 (1,175,400)	167 (962,428)	381 (1,718,818)	-
B. ambifaria **Unique to \ vs**	**AMMD**	**MC40-6**		
AMMD	-	190 (1,120,124)	-	-
MC40-6	193 (1,244,038)	-	-	-
B. mallei **Unique to \ vs**	**ATCC 23344**	**NCTC 10229**	**NCTC 10247**	**SAVP1**
ATCC 23344	-	20 (170,941)	13 (84,924)	38 (666,140)
NCTC 10229	21 (82,763)	-	0 (0)	44 (608,110)
NCTC 10247	22 (84,312)	8 (87,233)	-	46 (611,517)
SAVP1	7 (66,131)	14 (95,740)	9 (11,619)	-
B. pseudomallei **Unique to \ vs**	**1106a**	**1710b**	**668**	**K96243**
1106a	-	45 (127,005)	99 (351,322)	65 (190,015)
1710b	41 (332,186)	-	112 (547,644)	59 (345,501)
668	78 (311,275)	88 (296,951)	-	81 (316,064)
K96243	67 (340,747)	66 (292,446)	97 (513,906)	-

*Number of gaps > 500bp present in all pairwise strain comparisons (total bp in these gaps)

When we calculated the regions that were strain-specific (ie. not shared with another strain within the same species), up to 15% of the genes within a given genome were unique compared to the other available genomes. Interestingly, the unique genes were not equally distributed among the chromosomes (Figure 3.3), highlighting both the genomic versatility and the specialization of replicons within the *Burkholderia*. In addition, not all species display the same patterns, indicating that lifestyle and evolutionary trajectory (*B. mallei* for example has been undergoing substantial genome reduction), may play an important role. While at most 1% of the genes within any *B. mallei* strain were strain-specific, substantial fractions of the replicons in other species were accounted for by strain-specific or unique genes. For *B. pseudomallei*, depending on the strain, either of its two chromosomes carried up to 7% genic regions that are strain-specific. In contrast, the trend in the *Bcc* species is that the secondary chromosomes and plasmids carry more strain-specific genes as a function of the fraction of the replicon. Interestingly ~20-45% of the genes in chromosome 3 are strain-specific, while ~5-25%, and ~8-17%, of the genes on chromosome 2 and chromosome 1, respectively, fall into the strain-specific category (Figure 3.3). The explanation for the disparity in the two strains AU1054 and HI2424, is that they are highly similar and belong to the same clonal lineage (likely very recently diverged), and thus harbor similar genetic elements that are unique compared with the other two strains (see below). This highlights the need to consider both the methods and the subjects of analysis before interpreting results from broad-scale comparative genomics.

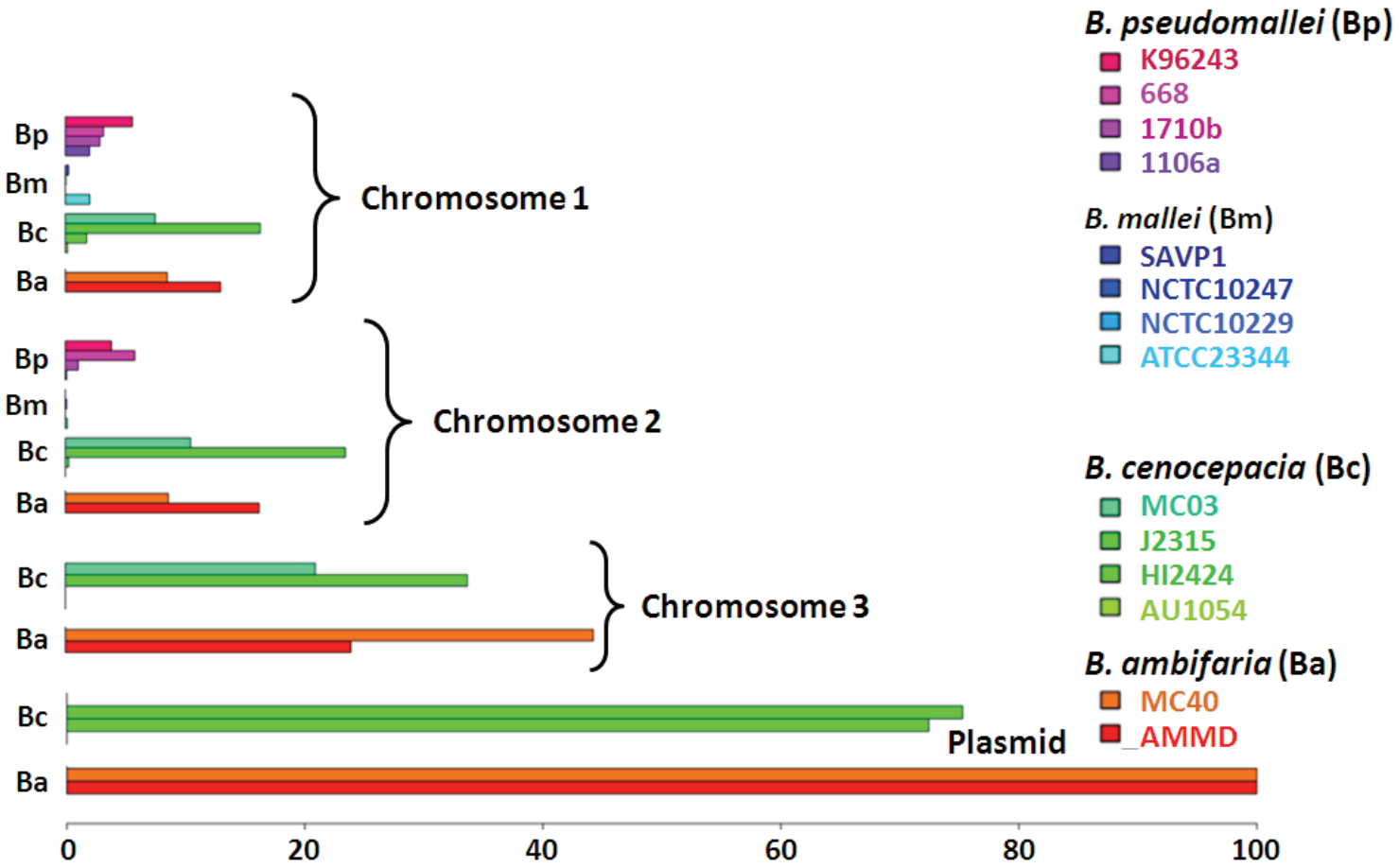


Figure 3.3. Distribution of strain-specific genes among the genomes of *Burkholderia* species. The numbers of genes unique to each strain were sorted by replicon, and the relative proportions of these strain-specific genes on each chromosome or plasmid are shown in percent of the replicon's genes. While the plasmids are mostly if not completely unique, there is also a substantial fraction of the chromosomal genes that are strain-specific.

Although genetically highly similar and indeed strains of the same *B. cenocepacia* clonal lineage, HI2424 and AU1054 may be specialized for different environments. Strain HI2424 was isolated from soil, while AU1054 is a clinical isolate from the blood of a patient with cystic fibrosis, therefore niche-specific differences may be anticipated. Further, we have previously shown that they respond differently when grown in either media mimicking cystic fibrosis sputum or in soil extract (90). Interestingly, among the few genes unique to AU1054 compared with the other *B. cenocepacia* isolates, is a cluster of three genes which include a transcriptional regulator, a methyltransferase, and an alkylhydroperoxidase D homolog. This latter family of enzymes has been implicated in persistence within macrophages (22), a common target for invasion by *Bc*c organisms (39), and appears to contribute to *Streptococcus pneumonia* virulence (61). In contrast, HI2424 carries 166 genes not present in AU1054, including >30 transcriptional regulators, perhaps explaining their differences in transcriptional response (90).

Although HI2424 contains a larger number of genes (176) not found in any of the other *B. cenocepacia* isolates, the majority appear to be of foreign origin, encoding many phage-related proteins, conjugation proteins and hypothetical proteins. Because HI2424 and AU1054 belong to the same lineage, they share many genes (391) not found in the other two isolates. Although many of these genes are also of foreign origin, one genomic island carries capsular polysaccharide biosynthesis, decoration, and transport capabilities (Bcen2424_0769- Bcen2424_0883). Such polysaccharides have been shown to be involved in virulence in both *B. mallei* and *B. pseudomallei* (17, 64, 69), and to be acquired by non-pathogenic *Burkholderia* isolates (70). These two strains also uniquely harbor a number of

putative antibiotic resistance genes, some of which appear to have been imported by phage.

Epidemic strain J2315 and environmental isolate MCO-3 carry 1481 and 753 strain-specific genes, respectively. While most of the genomic islands in the epidemic strain J2315 are of foreign origin and have already been addressed elsewhere (36), those in the maize rhizosphere strain MCO-3 have not been characterized. Interestingly, along with phage DNA, this strain harbors a number of regions, including a large genomic island carrying 172 genes that may play a role in survival in the rhizosphere, including a number of catabolic functions, such as those for aromatic ring cleavage (e.g. several dioxygenases) that may be involved in the breakdown of plant exudates or environmental contaminants. MCO-3 also harbors a >70 gene cluster encoding flagellar biosynthesis proteins, and another >50 gene island that carries unique Type III secretion system proteins, including a homolog of the host plant recognition and pathogenicity protein HrpB2 (68).

The *B. ambifaria* species has fewer members fully sequenced, thus these analyses are rather preliminary, however a number of interesting differences between isolates was observed. While both strains harbored a number of unique genes involved in sensing or interacting with external factors and responding with many unique transcriptional regulators, approximately 1/3 of the >1000 strain-specific genes are classified as hypothetical, with unknown function. Strain AMMD encodes a number of fatty acid biosynthesis proteins, while an abundance of transporters (including many sugar transporters), hemagluttinin proteins and >20 fimbrial proteins, are found in MC-40-6. While only a few Type I and II secretion system proteins were found to be AMMD-specific, a

larger number of unique Type I-IV secretion system proteins and >35 unique Type VI secretion system proteins were present in strain MC-40-6.

In *B. mallei*, all strains are believed to be of the same clonal lineage derived from a strain of *B. pseudomallei*, thus it was expected that this "species" would not be too diverse, and the largest number of regions unique to any one strain in two way comparisons was under 50 (see Table 3.2). Although the avirulent SAVP strain does carry a cluster of 7 genes, either fully or partially missing from the other strains (including FimA fimbriae), it has a number of large strain-specific deletions, including a >220 gene cluster that includes many type III secretion system proteins, which can be attributed to the attenuation of this strain. ATCC 23344 has 64 unique genes compared with other *B. mallei* strains, 61 of which are found in one island that appears to have been deleted from the other strains, and encodes a number of heavy metal transport systems. Few if any unique genes were found in either of the other two virulent strains of *B. mallei*.

All four *B. pseudomallei* were isolated from patients, and each carries between 80 and 310 genes that are not present in the other three. Interestingly, although the role of insects in transmission of the disease is uncertain (13), one of the large islands unique to 1710b carries a number of proteins that encode various components of insecticidal toxin complexes, suggesting the possible role of insects in the lifecycle of *B. pseudomallei*. The remaining regions unique to 1710b as well as all of the unique regions in 1106a appear to be of phage origin. The majority of the other two strains also appear to have been laterally acquired. While strain 668 does harbor a large 75 gene cluster encoding flagellar biosynthesis and chemotaxis functions that are not found in the other *B. pseudomallei*, a 25 gene cluster in K96243 encodes a number of metabolic proteins, and instead of having

been recently acquired via phage, may have been deleted from the other *B. pseudomallei* isolates.

Chromosome evolution

In addition to identifying all regions specific to any given strain, detailed nucleotide comparisons revealed anywhere between ~200 and ~265,000 SNPs, along with ~400 to ~30,000 small (1-10bp) insertions or deletions (indels) between strains of the same species (Table 3.3). Given the large number of mutations in all species aside from *B. mallei*, virtually every gene within the genome was affected by nsSNPs and/or indels. The number of SNPs appears to track very well with the phylogeny based on 31 housekeeping genes (Figure 3.1). For example, a larger number of SNPs within *B. cenocepacia* occur between the most distantly diverged J2315 strain and the other isolates, followed by MCO-3 and the other two closely related strains AU1054 and HI2424. The number of small indels appears to follow a similar pattern. Given the phylogeny of housekeeping genes, it is perhaps not surprising to see the smaller number of SNPs between *B. mallei* strains, and to a lesser extent, *B. pseudomallei*.

Table 3.3. SNPs and indels from pairwise genome comparisons among species.

B. cenocepacia Indels \ SNPs	AU1054	HI2424	J2315	MC0-3
AU1054	-	638	253,972	111,247
HI2424	398	-	263,298	114,032
J2315	29,892	30,449	-	265,047
MC0-3	12,481	12,544	30,115	-
B. ambifaria **Indels \ SNPs**	**AMMD**	**MC40-6**	**-**	**-**
AMMD	-	154,777	-	-
MC40-6	22,628	-	-	-
B. mallei **Indels \ SNPs**	**ATCC 23344**	**NCTC 10229**	**NCTC 10247**	**SAVP1**
ATCC 23344	-	959	904	626
NCTC 10229	4,220	-	205	737
NCTC 10247	4,474	1,317	-	710
SAVP1	3,185	3,549	3,832	-
B. pseudomallei **Indels \ SNPs**	**1106a**	**1710b**	**668**	**K96243**
1106a	-	20,128	33,432	21,775
1710b	17,681	-	33,572	19,963
668	25,255	25,283	-	33,350
K96243	19,221	18,722	26,374	-

*The number of SNPs (top right half, blue) and indels (bottom left half, red) found between any two strains of a species.

It is interesting to note that within a species, the ratio of SNPs to indels appears to be stable with the exception of AU1054 and HI2424, which are recently derived isolates of the same lineage of *B. cenocepacia*. Although for the latter pair a recent divergence hardly explains this phenomenon, the consistency of SNP:indel ratios between all other strains is congruent with the assumption of relatively stable mutation rates and efficacy of DNA

repair machinery. Perhaps more intriguing is the fact that only in *B. mallei* are there more small indels than SNPs (Figure 3.4). It is interesting to note that in *B. mallei*, there are also many rearrangements that occur between strains, despite their recent evolution from a common ancestor. It has been documented that SNPs outnumber indels even in humans (indels of up to 10kb were counted), which match the results of three of the species comparisons (55). However, to our knowledge this is the first report where indels are consistently observed to be more abundant than SNPs. A number of factors could be responsible, including fidelity of polymerase, mismatch repair, local DNA composition or genome structure and stability, etc., however further study will be required to understand the molecular basis of this selective mutation type. This may also be related to the fact that *B. mallei* is locked in an evolutionary trajectory of genome degradation, which could impact the rates and accumulation of particular mutations (abundance in indels and in genomic rearrangements), highlighting different modes of evolution apparent within one genus.

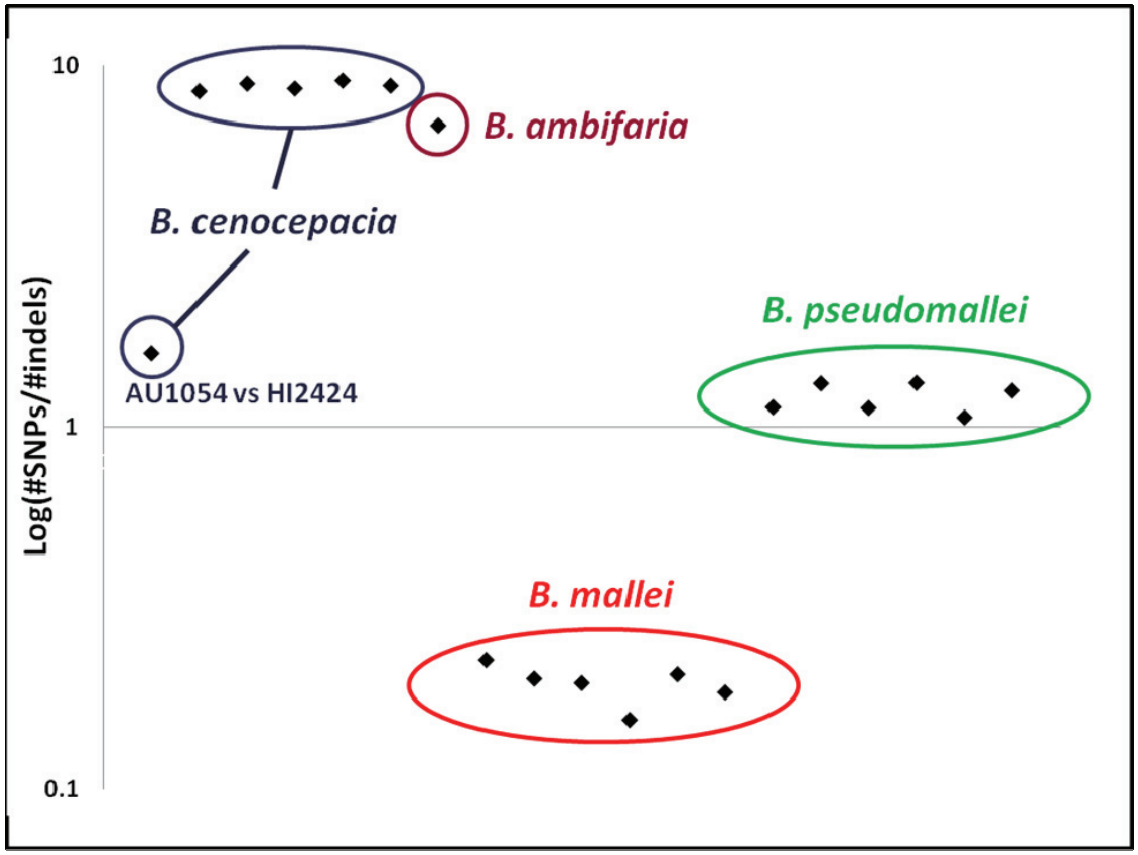


Figure 3.4. Ratio of SNPs to small indels among all strains of four *Burkholderia* species. The log of the ratio of SNPs to indels within each genome is shown. Each species clusters vertically along this graph with only *B. mallei* having fewer SNPs than indels. The two *Bcc* species have many more SNPs than indels with the exception of the recently diverged strains AU1054 and HI2424.

Given these interesting differences in mutation patterns among the species, we also investigated whether the different replicons varied with respect to rates of both synonymous (dS) and non-synonymous (dN) evolution, as has previously been observed already among *Burkholderia* species as well as within *B. pseudomallei* and *B. cenocepacia*

(10, 12, 56). The codons of all SNPs found to occur within genes in any of the strains of the four species were extracted, sorted by chromosome, and subjected to evolutionary analyses. The rates of evolution for *B. ambifaria* were estimated to have reached saturation, thus the results are inconclusive. Confirming the results from other studies however, the *B. pseudomallei* dS rates for both chromosomes were similar, but in *B. cenocepacia* the rate was elevated in chromosome 2 compared with chromosome 1, indicating comparable and elevated levels of synonymous substitution in the secondary chromosomes of these species. The dN rate in both species was confirmed to be elevated in secondary chromosomes compared with chromosome 1. Interestingly, in *B. mallei*, both dN and dS were found to be lower, indicating an decreased rate of evolution, in chromosome 2 compared with chromosome 1. The reason for this shift is unclear but may be linked to the smaller SNP:indel ratio noted above.

Comparative pangenomic analysis reveals the true "species" core

While the above analyses allow the comparison of trends between the four interrogated species, they do not directly compare the pangenomes. In order to compare all four species, a protein-based analysis was undertaken due to the divergence in sequence among all four species. Interspecies orthologs were determined by reciprocal best blast analysis that allowed the merging of protein families. While the pangenome analyses above highlighted the unique portions of each genome within a species, to address the species core, we compared the core genes from all four species together (Figure 3.5). This analysis left from 171-832 genes that were specific to the core of a species, as defined by not being found in the core of another species. To narrow down the true genic definition of

these *Burkholderia* species, the remaining set of core genes were then compared with the pangenome of all the other species. This resulted in an addition ~50% reduction in the number of core genes present within a given species and absent from the pangenomes of the three other species (Figure 3.5). These 64 *B. mallei*, 221 *B. pseudomallei*, 350 *B. cenocepacia*, and 359 *B. ambifaria* genes are thus species-specific core genes that may be responsible for species-specific traits. Closer inspection of these genes by functional COG category revealed a somewhat random distribution of functions with a relatively large fraction of the unique core genomes for *B. cenocepacia* and *B. ambifaria* assigned to proteins with unknown function (Figure 3.5). Illustrating again the difference between *B. mallei* and the other species, in terms of distribution of genes among the replicons, *B. mallei* is the only species that carries more of its species-specific genes on chromosome 1 (Figure 3.5). The majority of these species-specific genes are found on chromosome 2, and in the case of *B. ambifaria* and *B. cenocepacia*, chromosome 3 also harbors more of the species-specific genes that does chromosome 1.

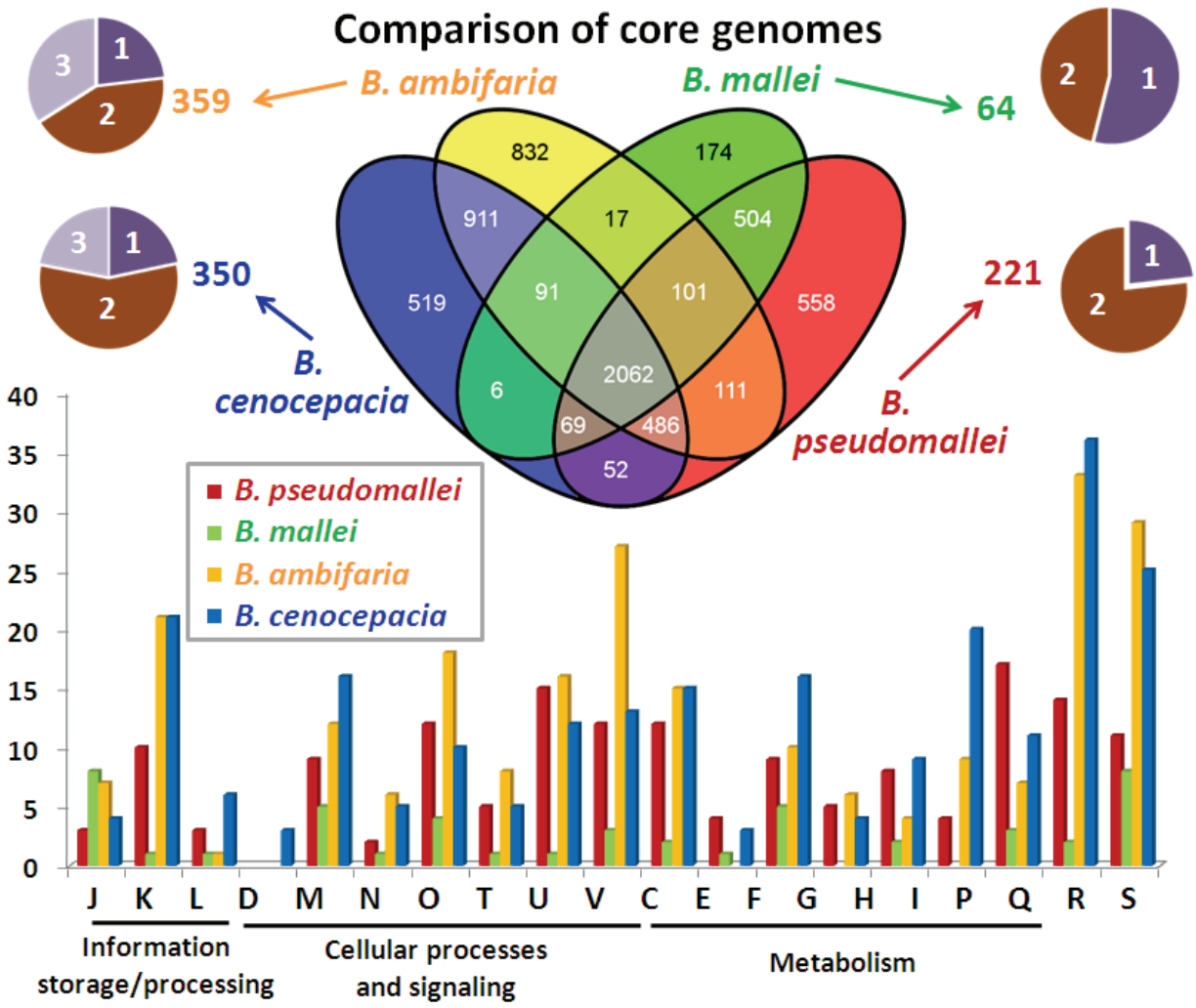


Figure 3.5. Comparative pangenome analyses. The Venn diagram displays the comparison of species core genomes. The numbers at the tips of the Venn represent species-specific core genes not present in other species cores. When further subtracting those genes found within the pangenomes of the other species, the number of species-specific genes decreases again (arrows with numbers). The pie charts display the distribution of these genes among the chromosomes present in the species. These same remaining genes are broken down by COG functional categories below. The COG categories are as follows, J: Translation, ribosomal structure and biogenesis; K: Transcription; L: Replication, recombination and repair; D: Cell cycle control, cell division, chromosome

partitioning; T: Signal transduction mechanisms; V: Defense mechanisms; W: Extracellular structures; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Posttranslational modification, protein turnover, chaperones; U: Intracellular trafficking, secretion, and vesicular transport; C: Energy production and conversion; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R: General function prediction only; S: Unknown.

Conclusions and implications when studying bacterial "species"

All the methods presented here can be broadly applied to bacterial (or archaeal) groups of phylogenetically closely related strains. However, it is important to keep in mind the precise relatedness of strains, their life histories, their potential for molecular diversification, etc. before interpretation and generalization of resulting data. It is perhaps obvious that not all named species are equivalent in terms of phylogenetic diversity, thus rules for one may not necessarily be applicable to the other. Although this is perhaps only a semantic issue, it has been hotly debated, as has the very definition and even existence of bacterial species or whether only a genomic continuum exists (e.g. (20, 42)). Research within one group has led to both to the conclusion that genomic diversity in accessory genes can exist within a highly coherent species genomic backbone, as well as question whether the term 'species' could even be applied to a group of organisms undergoing constant homogenization via rampant lateral gene transfer (49, 60). However, given that

for some groups of bacteria, certain phenotypes are viewed to be perhaps more distinctive or important than others (e.g. virulent for humans) and sufficient to warrant a distinct (species) name, these are not necessarily the traits that are actively undergoing selective pressure or even of importance for the bacteria in its natural environment. Although in this review, we do not preoccupy ourselves with this debate, it is clear that further genomic comparisons of closely related groups, as well as metagenomic sampling and population genomics will help illuminate this topic further.

Here we present an analysis of the genomic material shared among strains of the same named species (of *Burkholderia*), along with analysis of portions missing from one or more strains. This has been termed pangenomics, at least when applied to conserved genes or proteins found within strains of the same species. The pangenome concept is not restricted to intraspecies analyses, and has been expanded to the genus and even the bacterial domain ((16, 48, 83), Chapter 2). These analyses have often been performed to better understand the variability of "gene space" and of metabolic potential within large groups of organisms, however some have also used gene content alone (ie. presence or absence) to derive evolutionary relationships (6, 83). While this has been shown to have merit in certain circumstances, this method of analysis is effectively the summation of several processes without well-defined evolutionary rates: deletion of genomic regions carrying genes, and insertion of DNA after acquisition via lateral transfer. Given that these molecular processes can affect many genes in a single event, and given the possibility of homologous recombination between isolates (58), this analysis may be prone to error in some instances. The use of indels and SNPs, particularly those not prone to selective

pressure, would be the best features used for phylogeny, however identification of these specific mutations is sometimes challenging.

While species core genes may be used to reconstruct phylogenetic relationships, given the gene flux and accelerated evolution of secondary chromosomes presented here in the case of *Burkholderia*, such an analysis may be misleading. Smaller gene subsets however, have been used to robustly delineate phylogenetic lineages within the bacterial tree of life (11, 50, 89), due in part to the choice of genes that are specifically not prone to lateral gene transfer or positive selection. We argue that efforts must take all of these factors into account before attempting to reconstruct and interpret multi-gene based organism phylogenies.

We present complementary protein-based as well as more detailed nucleotide-based methods for understanding the genetic variability within *Burkholderia* species. It is clear that the *Burkholderia* genome is fluid, and comparisons between named species indicate differences between species and between chromosomes within species and their prevalence for entertaining novel sequences, as well as differences between species in the abundance and types of mutations (indels and SNPs) found. With respect to rates of evolution, indels have received far less attention that SNPs or even large structural variations (rearrangements), in part due to the difficulty in detecting such mutations, as well as the challenge in ascribing evolutionary models to various types of indels (38, 55). As shown here, the rate of SNP accumulation versus that of indels is not stable between species, even if within a species, there appears to be some pattern to this ratio. Various factors may play a role in this observed ratio, including polymerase error, DNA repair or

other molecular mechanisms, environmental or lifestyle factors, genomic DNA signatures and base composition, etc. (91).

As has been shown before (10, 12) and confirmed here, the different chromosomes within *Burkholderia* appear to be under different selective constraints, which has important implications in evolutionary strategies, since secondary chromosomes may aid in 'sampling' genes, enable rapid evolution via sequence modification, and testing various gene combinations. In this sense, the variable, sometimes called "dispensable", genome may not be so variable given one particular habitat, niche or lifestyle. Pangenome and detailed intra-'species' comparisons of this group of bacteria highlight the importance of acknowledging that any genome is but a snapshot in the evolutionary trajectory of a given lineage, and that the bacterial genome is constantly under flux (in part dependent on environment, and the ability of some bacteria to access and uptake foreign material). Metagenomic and population analyses hold great promise in studying the accessory or flexible genome and will undoubtedly provide great insight into the genomic diversity and mechanisms underlying the diversification of bacterial genomes.

References

1. **Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz.** 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature **430:**551-4.

2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-10.

3. **Baldwin, A., E. Mahenthiralingam, K. M. Thickett, D. Honeybourne, M. C. Maiden, J. R. Govan, D. P. Speert, J. J. Lipuma, P. Vandamme, and C. G. Dowson.** 2005. Multilocus sequence typing scheme that provides both species and strain differentiation for the Burkholderia cepacia complex. J Clin Microbiol **43:**4665-73.

4. **Beiko, R. G., and N. Hamilton.** 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol Biol **6:**15.

5. **Bergthorsson, U., and H. Ochman.** 1998. Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol Biol Evol **15:**6-16.

6. **Bohlin, J., L. Snipen, A. Cloeckaert, K. Lagesen, D. Ussery, A. B. Kristoffersen, and J. Godfroid.** 2010. Genomic comparisons of Brucella spp. and closely related bacteria using base compositional and proteome based methods. BMC Evol Biol **10:**249.

7. **Bottacini, F., D. Medini, A. Pavesi, F. Turroni, E. Foroni, D. Riley, V. Giubellini, H. Tettelin, D. van Sinderen, and M. Ventura.** 2010. Comparative genomics of the genus Bifidobacterium. Microbiology **156:**3243-54.

8. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis Comparison Tool. Bioinformatics **21:**3422-3.

9. **Cesarini, S., A. Bevivino, S. Tabacchioni, L. Chiarini, and C. Dalmastri.** 2009. RecA gene sequence and Multilocus Sequence Typing for species-level resolution of Burkholderia cepacia complex isolates. Lett Appl Microbiol **49:**580-8.

10. **Chain, P. S., V. J. Denef, K. T. Konstantinidis, L. M. Vergez, L. Agullo, V. L. Reyes, L. Hauser, M. Cordova, L. Gomez, M. Gonzalez, M. Land, V. Lao, F. Larimer, J. J. LiPuma, E. Mahenthiralingam, S. A. Malfatti, C. J. Marx, J. J. Parnell, A. Ramette, P. Richardson, M. Seeger, D. Smith, T. Spilker, W. J. Sul, T. V. Tsoi, L. E. Ulrich, I. B. Zhulin, and J. M. Tiedje.** 2006. Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A **103:**15280-7.

11. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283-7.

12. **Cooper, V. S., S. H. Vohr, S. C. Wrocklage, and P. J. Hatcher.** 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput Biol **6:**e1000732.

13. **Dance, D. A.** 2000. Ecology of Burkholderia pseudomallei and the interactions between environmental Burkholderia spp. and human-animal hosts. Acta Trop **74:**159-68.

14. **Daubin, V., M. Gouy, and G. Perriere.** 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res **12:**1080-90.

15. **DeLong, E. F., and N. R. Pace.** 2001. Environmental diversity of bacteria and archaea. Syst Biol **50:**470-8.

16. **den Bakker, H. C., C. A. Cummings, V. Ferreira, P. Vatta, R. H. Orsi, L. Degoricija, M. Barker, O. Petrauskene, M. R. Furtado, and M. Wiedmann.** 2010. Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics **11:**688.

17. **DeShazer, D., D. M. Waag, D. L. Fritz, and D. E. Woods.** 2001. Identification of a Burkholderia mallei polysaccharide gene cluster by subtractive hybridization and demonstration that the encoded capsule is an essential virulence determinant. Microb Pathog **30:**253-69.

18. **Doolittle, W. F.** 1999. Phylogenetic classification and the universal tree. Science **284:**2124-9.

19. **Doolittle, W. F., and R. T. Papke.** 2006. Genomics and the bacterial species problem. Genome Biol **7:**116.

20. **Doolittle, W. F., and O. Zhaxybayeva.** 2009. On the origin of prokaryotic species. Genome Res **19:**744-56.

21. **Edgar, R. C.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:**113.

22. **Farivar, T. N., P. J. Varnousfaderani, and A. Borji.** 2008. Mutation in alkylhydroperoxidase D gene dramatically decreases persistence of Mycobacterium bovis bacillus calmette-guerin in infected macrophage. Indian J Med Sci **62:**275-82.

23. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al.** 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:**496-512.

24. **Fraser-Liggett, C. M.** 2005. Insights on biology and evolution from microbial genome sequencing. Genome Res **15:**1603-10.

25. **Fulton, D. L., Y. Y. Li, M. R. Laird, B. G. Horsman, F. M. Roche, and F. S. Brinkman.** 2006. Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics **7:**270.

26. **Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings.** 2005. Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol **3:**733-9.

27. **Godoy, D., G. Randle, A. J. Simpson, D. M. Aanensen, T. L. Pitt, R. Kinoshita, and B. G. Spratt.** 2003. Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, Burkholderia pseudomallei and Burkholderia mallei. J Clin Microbiol **41:**2068-79.

28. **Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence.** 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol **19:**2226-38.

29. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696-704.

30. **Guo, F. B., L. W. Ning, J. Huang, H. Lin, and H. X. Zhang.** 2010. Chromosome translocation and its consequence in the genome of Burkholderia cenocepacia AU-1054. Biochem Biophys Res Commun **403:**375-9.

31. **Hall, B. G., G. D. Ehrlich, and F. Z. Hu.** 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology **156:**1060-8.

32. **Hanage, W. P., C. Fraser, and B. G. Spratt.** 2005. Fuzzy species among recombinogenic bacteria. BMC Biol **3:**6.

33. **Hanage, W. P., C. Fraser, and B. G. Spratt.** 2006. The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol **239:**210-9.

34. **Hey, J.** 2001. The mind of the species problem. Trends Ecol Evol **16:**326-329.

35. **Hiller, N. L., B. Janto, J. S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N. E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S. D. Bentley, J. C. Post, G. D. Ehrlich, and F. Z. Hu.** 2007. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol **189:**8186-95.

36. **Holden, M. T., H. M. Seth-Smith, L. C. Crossman, M. Sebaihia, S. D. Bentley, A. M. Cerdeno-Tarraga, N. R. Thomson, N. Bason, M. A. Quail, S. Sharp, I. Cherevach, C. Churcher, I. Goodhead, H. Hauser, N. Holroyd, K. Mungall, P. Scott, D. Walker, B. White, H. Rose, P. Iversen, D. Mil-Homens, E. P. Rocha, A. M. Fialho, A. Baldwin, C. Dowson, B. G. Barrell, J. R. Govan, P. Vandamme, C. A. Hart, E. Mahenthiralingam, and J. Parkhill.** 2009. The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol **191:**261-77.

37. **Holden, M. T., R. W. Titball, S. J. Peacock, A. M. Cerdeno-Tarraga, T. Atkins, L. C. Crossman, T. Pitt, C. Churcher, K. Mungall, S. D. Bentley, M. Sebaihia, N. R. Thomson, N. Bason, I. R. Beacham, K. Brooks, K. A. Brown, N. F. Brown, G. L. Challis, I. Cherevach, T. Chillingworth, A. Cronin, B. Crossett, P. Davis, D. DeShazer, T. Feltwell, A. Fraser, Z. Hance, H. Hauser, S. Holroyd, K. Jagels, K. E. Keith, M. Maddison, S. Moule, C. Price, M. A. Quail, E. Rabbinowitsch, K. Rutherford, M. Sanders, M. Simmonds, S. Songsivilai, K. Stevens, S. Tumapa, M. Vesaratchavest, S. Whitehead, C. Yeats, B. G. Barrell, P. C. Oyston, and J. Parkhill.** 2004. Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proc Natl Acad Sci U S A **101:**14240-5.

38. **Hollister, J. D., J. Ross-Ibarra, and B. S. Gaut.** 2010. Indel-associated mutation rate varies with mating system in flowering plants. Mol Biol Evol **27:**409-16.

39. **Huynh, K. K., J. D. Plumb, G. P. Downey, M. A. Valvano, and S. Grinstein.** 2010. Inactivation of macrophage Rab7 by Burkholderia cenocepacia. J Innate Immun **2:**522-33.

40. **Jain, R., M. C. Rivera, and J. A. Lake.** 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A **96:**3801-6.

41. **Keswani, J., and W. B. Whitman.** 2001. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. Int J Syst Evol Microbiol **51:**667-78.

42. **Konstantinidis, K. T., A. Ramette, and J. M. Tiedje.** 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci **361:**1929-40.

43. **Konstantinidis, K. T., and J. M. Tiedje.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A **102:**2567-72.

44. **Korber, B.** 2000. Computational analysis of HIV molecular sequences, p. 55-72 *In* A. Rodrigo and H. L. Gerald (ed.), HIV Signature and Sequence Variation Analysis. Kluwer Academic Publishers Dordrecht, Netherlands.

45. **Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg.** 2004. Versatile and open software for comparing large genomes. Genome Biol **5:**R12.

46. **Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas, and V. P. Gannon.** 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics **11:**461.

47. **Lan, R., and P. R. Reeves.** 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol **8:**396-401.

48. **Lapierre, P., and J. P. Gogarten.** 2009. Estimating the size of the bacterial pan-genome. Trends Genet **25:**107-10.

49. **Legault, B. A., A. Lopez-Lopez, J. C. Alba-Casado, W. F. Doolittle, H. Bolhuis, F. Rodriguez-Valera, and R. T. Papke.** 2006. Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics **7:**171.

50. **Lerat, E., V. Daubin, and N. A. Moran.** 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol **1:**E19.

51. **Lessie, T. G., W. Hendrickson, B. D. Manning, and R. Devereux.** 1996. Genomic complexity and plasticity of Burkholderia cepacia. FEMS Microbiol Lett **144:**117-28.

52. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A **95:**3140-5.

53. **Medini, D., D. Serruto, J. Parkhill, D. A. Relman, C. Donati, R. Moxon, S. Falkow, and R. Rappuoli.** 2008. Microbiology in the post-genomic era. Nat Rev Microbiol **6:**419-30.

54. **Mira, A., A. B. Martin-Cuadrado, G. D'Auria, and F. Rodriguez-Valera.** 2010. The bacterial pan-genome:a new paradigm in microbiology. Int Microbiol **13:**45-57.

55. **Mullaney, J. M., R. E. Mills, W. S. Pittard, and S. E. Devine.** 2010. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet **19:**R131-6.

56. **Nandi, T., C. Ong, A. P. Singh, J. Boddey, T. Atkins, M. Sarkar-Tyson, A. E. Essex-Lopresti, H. H. Chua, T. Pearson, J. F. Kreisberg, C. Nilsson, P. Ariyaratne, C. Ronning, L. Losada, Y. Ruan, W. K. Sung, D. Woods, R. W. Titball, I. Beacham, I. Peak, P. Keim, W. C. Nierman, and P. Tan.** 2010. A genomic survey of positive selection in Burkholderia pseudomallei provides insights into the evolution of accidental virulence. PLoS Pathog **6:**e1000845.

57. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3:**418-26.

58. **Ochman, H., E. Lerat, and V. Daubin.** 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A **102 Suppl 1:**6595-9.

59. **Ota, T., and M. Nei.** 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol Biol Evol **11:**613-9.

60. **Papke, R. T., O. Zhaxybayeva, E. J. Feil, K. Sommerfeld, D. Muise, and W. F. Doolittle.** 2007. Searching for species in haloarchaea. Proc Natl Acad Sci U S A **104:**14092-7.

61. **Paterson, G. K., C. E. Blue, and T. J. Mitchell.** 2006. An operon in Streptococcus pneumoniae containing a putative alkylhydroperoxidase D homologue contributes to virulence and the response to oxidative stress. Microb Pathog **40:**152-60.

62. **Payne, G. W., P. Vandamme, S. H. Morgan, J. J. Lipuma, T. Coenye, A. J. Weightman, T. H. Jones, and E. Mahenthiralingam.** 2005. Development of a recA gene-based identification approach for the entire Burkholderia genus. Appl Environ Microbiol **71:**3917-27.

63. **Rasko, D. A., M. J. Rosovitz, G. S. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, and J. Ravel.** 2008. The pangenome structure of Escherichia coli:

comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol **190:**6881-93.

64. **Reckseidler-Zenteno, S. L., R. DeVinney, and D. E. Woods.** 2005. The capsular polysaccharide of Burkholderia pseudomallei contributes to survival in serum by reducing complement factor C3b deposition. Infect Immun **73:**1106-15.

65. **Rocha, E. P.** 2004. Order and disorder in bacterial genomes. Curr Opin Microbiol **7:**519-27.

66. **Rocha, E. P.** 2004. The replication-related organization of bacterial genomes. Microbiology **150:**1609-27.

67. **Rossello-Mora, R., and R. Amann.** 2001. The species concept for prokaryotes. FEMS Microbiol Rev **25:**39-67.

68. **Rossier, O., G. Van den Ackerveken, and U. Bonas.** 2000. HrpB2 and HrpF from Xanthomonas are type III-secreted proteins and essential for pathogenicity and recognition by the host plant. Mol Microbiol **38:**828-38.

69. **Sarkar-Tyson, M., J. E. Thwaite, S. V. Harding, S. J. Smither, P. C. Oyston, T. P. Atkins, and R. W. Titball.** 2007. Polysaccharides and virulence of Burkholderia pseudomallei. J Med Microbiol **56:**1005-10.

70. **Sim, B. M., N. Chantratita, W. F. Ooi, T. Nandi, R. Tewhey, V. Wuthiekanun, J. Thaipadungpanit, S. Tumapa, P. Ariyaratne, W. K. Sung, X. H. Sem, H. H. Chua, K. Ramnarayanan, C. H. Lin, Y. Liu, E. J. Feil, M. B. Glass, G. Tan, S. J. Peacock, and P. Tan.** 2010. Genomic acquisition of a capsular polysaccharide virulence cluster by non-pathogenic Burkholderia isolates. Genome Biol **11:**R89.

71. **Stackebrandt, E., and J. Ebers.** 2006. Taxonomic parameters revisited: tarnished gold standards, p. 152-155, Microbiol Today, vol. 33.

72. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kampfer, M. C. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52:**1043-7.

73. **Stackebrandt, E., and B. M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol **44:**846-849.

74. **Staley, J. T.** 2006. The bacterial species dilemma and the genomic-phylogenetic species concept. Philos Trans R Soc Lond B Biol Sci **361:**1899-909.

75. **Staley, J. T.** 1997. Biodiversity: are microbial species threatened? Curr Opin Biotechnol **8:**340-5.

76. **Staley, J. T.** 2009. Universal species concept: pipe dream or a step toward unifying biology? J Ind Microbiol Biotechnol **36:**1331-6.

77. **Talavera, G., and J. Castresana.** 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol **56:**564-77.

78. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

79. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser.** 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A **102:**13950-5.

80. **Thompson, C. C., A. C. Vicente, R. C. Souza, A. T. Vasconcelos, T. Vesth, N. Alves, Jr., D. W. Ussery, T. Iida, and F. L. Thompson.** 2009. Genomic taxonomy of Vibrios. BMC Evol Biol **9:**258.

81. **Thompson, J. D., T. J. Gibson, and D. G. Higgins.** 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics **Chapter 2:**Unit 2 3.

82. **Trost, B., M. Haakensen, V. Pittet, B. Ziola, and A. Kusalik.** 2010. Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. BMC Microbiol **10:**258.

83. **Ussery, D. W., K. Kiil, K. Lagesen, T. Sicheritz-Ponten, J. Bohlin, and T. M. Wassenaar.** 2009. The genus burkholderia: analysis of 56 genomic sequences. Genome Dyn **6:**140-57.

84. **Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings.** 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev **60:**407-38.

85. **Ward, D. M.** 1998. A natural species concept for prokaryotes. Curr Opin Microbiol **1:**271-7.

86. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, L. Krichevsky, L. H. Moore, W. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Truper.** 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol **37:**463-464.

87. **Welch, R. A., V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A **99:**17020-4.

88. **Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin.** 2002. Genome trees and the tree of life. Trends Genet **18:**472-9.

89. **Wu, M., and J. A. Eisen.** 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol **9:**R151.

90. **Yoder-Himes, D. R., P. S. Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, J. M. Tiedje, and R. Sorek.** 2009. Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. Proc Natl Acad Sci U S A **106:**3976-81.

91. **Zanders, S., X. Ma, A. Roychoudhury, R. D. Hernandez, A. Demogines, B. Barker, Z. Gu, C. D. Bustamante, and E. Alani.** 2010. Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a Bayesian approach. Genetics **186:**493-503.

Chapter 4.
Circulating strains of an epidemic clonal lineage (ET-12) of *Burkholderia cenocepacia*: Genomic variation in clinical isolates over two decades.

Introduction

The *Burkholderia cepacia* complex (*Bcc*) is a group of very closely related betaproteobacteria comprised of a growing number of species (current count is at 17) that, although genetically similar, have a great metabolic versatility to inhabit a number of different environments, including soil, surface waters, and human and plant tissues (9, 28-30, 42-43). The *Bcc* encompasses both beneficial plant rhizosphere symbionts as well as the more sinister agents of onion rot and opportunistic pathogens of the human lung. It has been found that some species appear to harbor greater capacity to inhabit certain environmental niches than others, as some species and subspecies are recovered primarily from soil, whereas others are most often found associated with eukaryotic hosts (3, 8-9, 30).

Two members of the *Bcc* have captured the attention of clinicians in recent decades, *B. cenocepacia* and to a lesser extent *B. multivorans*, because they are the dominant isolates recovered from the immunocompromised lungs of cystic fibrosis (CF) patients. *B. cenocepacia* gained notoriety due to its particularly devastating effects in patients with CF, known as "cepacia syndrome" which accelerates the decline of respiratory function, and ultimately results in death (22). Complicating the control of *B. cenocepacia* outbreaks within CF patient populations is the high degree of person-to-person transmissibility (15) and inherent antibiotic resistance of the bacterium via efflux pumps (11, 16). Other virulence factors that have been identified in some *B. cenocepacia* strains include quorum-sensing systems that regulate gene expression (26), siderophore synthesis enabling iron scavenging (2), exopolysaccharide production for protective capsule formation (14), and

outer membrane structures such as the cable pilus that confer the ability to adhere to host cells (36-37).

The phylogeny of *B. cenocepacia* strains based on *recA* sequences supports the subdivision of this species into 4 subgroups (41), with most of the strains implicated in human disease belonging to subgroups IIIA and IIIB, so named because *B. cenocepacia* was formerly called *Bcc* genomovar III. In the hopes of gaining insight into the virulence mechanisms of clinical *B. cenocepacia* strains that allow them to thrive in the human lung niche, we and others have been pursuing genome-based studies of this important group of pathogens (33, 45, 47). We have sequenced two distinct strains of the PHDC lineage, an epidemic IIIB lineage prominent in North American CF patients, that have been used as the basis for study of transcriptional responses under soil and CF sputum conditions (45). The 8.06 Mb genome sequence of IIIA strain J2315 was more recently published in 2009 (21). This strain belongs to the ET-12 (electrophoretic type 12) lineage implicated in epidemics in Canada and Europe and was isolated from the sputum of a CF patient in Edinburgh, UK in 1989 (15, 23, 41). To obtain insights into the genetic diversity and evolution of this lineage, we sought to sequence several strains isolated from patients during outbreaks among CF populations.

The speed with which next generation sequencing platforms can generate genomic sequences has made feasible the exploration of a number of strains in a single sequencing run over the course of only a week. Although these shorter reads have not yet replaced longer sequencing read approaches as the staple for whole genome sequencing, short read sequencing of strains that are very closely related to a reference genome is an amenable task. The sequencing depth allows reliable single nucleotide polymorphism (SNP)

observations as well as the detection of deletions compared to a reference sequence. Here, millions of Illumina sequence reads of four ET-12 lineage strains have been compared with the genome of the multidrug-resistant J2315 isolate. These new sequences and their comparison illustrate the utility of such high-throughput sequencing and provide insight into the evolutionary strategy of circulating epidemic ET-12 strains.

Materials and Methods

Strains and Sequencing

For this study, four strains of the ET-12 lineage of *B. cenocepacia* (Table 4.1) were selected for sequencing on the Illumina (formerly Solexa) GA II instrument. The strains J2315 (clinical isolate of lineage ET-12), AU1054 (clinical isolate of lineage PHDC), and HI2424 (agricultural soil isolate of lineage PHDC) were used as reference for all comparisons.

Table 4.1. Names and descriptions of strains selected for sequencing.

B. cenocepacia Strains	Year	Origin	Description
AU16956	1999	CF patient	Toronto
AU16958	2001	CF patient	Toronto
HI4277	2008	CF patient	Toronto, recent outbreak
HI4278	2008	CF patient	Toronto, recent outbreak
J2315	1989	CF patient	Edinburgh, UK
AU1054	1999	Blood of CF patient	Philadelphia
HI2424	1999	Onion field soil isolate	Upstate New York

Strains AU16956 and AU16958 were isolated during an earlier outbreak in a Toronto clinic, while the two other strains, HI4277 and HI4278, were isolated from cystic fibrosis patients in the fall of the 2008 Toronto outbreak. We selected these latter two strains because the respective patients had very different outcomes (one died quickly, while the other is still alive).

The four strains were each sequenced using a single channel (lane) of Illumina using a paired-end approach, 72 cycles each (resulting in 72bp paired-end reads), and following Illumina protocols. Between 21 and 24 million reads were generated for each of the four lanes of Illumina.

Illumina sequence quality evaluation

The raw Illumina sequence reads contained some of very poor quality. These were subjected to an automated filtering process via Illumina software (to pass filter) and underwent further read screening and trimming to a quality of score 20 (one error per 100bp) using a Perl script. Although the Illumina filtering software does not trim reads but removes reads with poor quality throughout, thus maintaining the length of the reads, our quality filtering/trimming protocol removes poor quality bases at the 3' end of reads and requires a good overall quality of the read. Reads that pass this method may be shorter in length than the raw reads (see Figure 4.1 and Table 4.2 for results).

Read-based analysis by mapping to a reference genome

The trimmed reads were mapped to the reference genomes using Mosaik 1.0.1388 (http://bioinformatics.bc.edu/marthlab/Mosaik) with default parameters for hash size (of 15), with up to12 mismatches allowed and a minimum alignment length of 35bp. These were chosen in part to allow easier comparison of trimmed and untrimmed reads. Mosaik was chosen for this analysis since it allows gapped alignments, reference-guided assembly, has other useful features for coverage calculation and can provide alignment outputs in the now standard SAM format such that they can be used interchangeably with other next-

generation sequence analysis tools. The aligned results were processed with SAMtools (v0.1.7; http://samtools.sourceforge.net/) for conversion to BAM format and for calling single nucleotide polymorphisms (SNPs). The resulting output was filtered to rule out error-prone variant calls (27).

Using a series of Perl and Python scripts, the coordinates of the potential SNPs and of missing "gap" locations were mapped to the NCBI annotation files of the reference in order to identify: 1) Gene names and product descriptions for those that contain potential SNPs; 2) Whether the SNP mutations result in non-synonymous or synonymous mutations (ie. result in a change or in no change in encoded amino acid; 3) Small insertions or deletions (indels) that may alter the function of genes; 4) Large deletions with respect to the reference and any genes that reside within these large gaps; and 5) The distribution of these SNPs and indels among the replicons (which chromosome or plasmid). To understand the possible functional constraints of SNP, indel and gap mutations, the functional classification of all affected genes were also calculated.

Analysis of genomes with assembly of reads

All assembly was done using *de novo* assembly techniques. Trimmed and untrimmed reads were assembled using Velvet with a range of Kmers between 41 and 53 (46). Contigs from each Kmer based assembly were then combined using Newbler (Life Technologies) and Minimus2 (AMOS; http://amos.sourceforge.net). The resulting contigs were assumed to be the best assembly possible from the available reads and were aligned to the reference sequence *B. cenocepacia* J2315 using NUCmer (part of the MUMmer system (24); http://mummer.sourceforge.net/) with default settings. Contigs that matched at an

identity > 99% and over at least 50% of the contig were used for subsequent SNP and gap analysis.

SNPs and indels were called using the show-snp tool (AMOS). Similar to the read-mapping analysis, the locations of gaps and nucleotide changes were analyzed using Perl scripts with annotation files to determine the genes with SNPs that resulted in synonymous or non-synonymous changes or if indels affected a protein. SNPs and indels in non-coding regions were not analyzed. Gap analysis was performed using alignment coordinates and determining the genes present in the reference but not covered by the aligned contigs. These results were compared with the read-mapping results using Perl scripts.

Regions within contigs that were larger than 500bp were considered novel sequences compared with the reference. Fasta sequences representing these regions were extracted and submitted to gene calling, blast analysis (vs nr and nt), and together with flanking regions were carefully aligned to the reference genome using TblastX.

Results

Sequencing four *B. cenocepacia* clinical isolates

The Illumina sequencing platform was used to quickly and inexpensively sample the genomes of four *B. cenocepacia* strains of the ET-12 lineage. Between 21 and 24 million reads were generated and >75% of these were maintained after all quality screening and filtering (Table 4.2 and Figure 4.1).

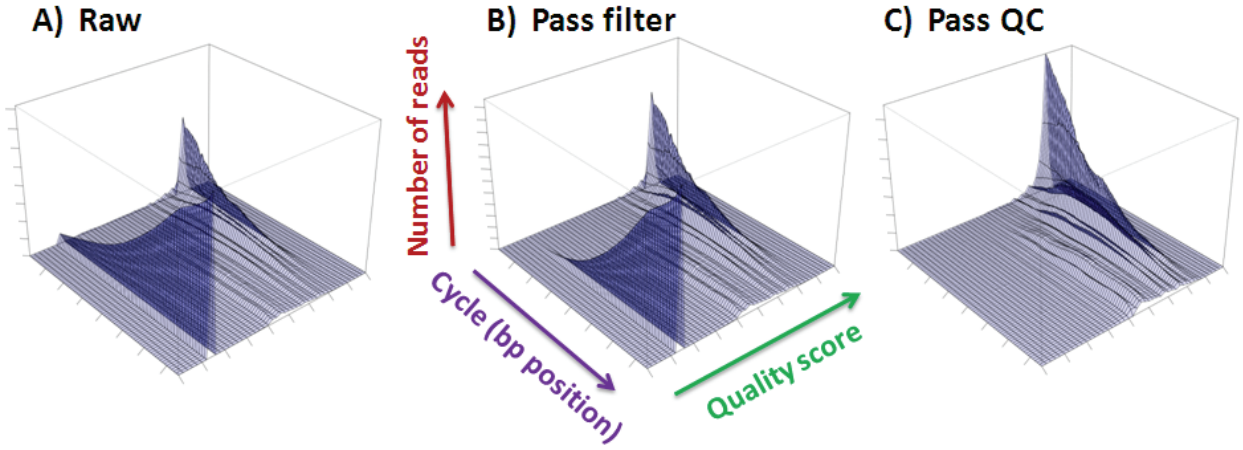


Figure 4.1. Effect of quality and trimming on Illumina reads. The per-base quality of all reads is displayed for the paired-end lane of HI4277 (the data for other strains appear highly similar, see Table 4.2 for statistics). The position of the base, or "cycle" within the read is displayed on the X axis, all reads are 76bp long. The Z axis shows the quality of the base, where 20 (error rate of approximately 1%) or greater are sufficient for assembly. The Y axis shows the number of reads, where at position X, the quality is Z. Most reads begin with high quality score bases, but A) Raw reads show the large fraction of bases that have very poor quality (near 0). B) Reads that pass the internal Illumina filtering algorithms, and shows that a number of very poor quality bases have been removed, primarily at the beginning of the read. C) Reads that pass the quality control (QC), which includes quality trimming of the data to remove all poor quality bases and allow more accurate and straightforward analysis.

Table 4.2. Sequencing statistics for four strains of the ET-12 *B. cenocepacia* lineage.

B. cenocepacia Strains	Number of raw 75bp reads	Number of pass filter 75bp reads	Number of pass QC* reads	QC* average read length (bp)
AU16956	21,841,221	18,573,095	17,301,011	55.8
AU16958	22,839,719	19,539,884	18,050,108	55.7
HI4277	24,170,663	20,055,054	18,545,887	57.9
HI4278	23,729,743	19,663,740	18,208,559	58.5

*Number of reads and bases that pass quality metrics as specified in Materials and Methods

Mapping reads to reference genomes reveals major differences between strains

In order to validate the sequence reads, their random distribution and their origin, they were mapped (aligned) to three available and complete *B. cenocepacia* genomes, including the ET-12 lineage highly epidemic strain J2315, and two PHDC lineage strains isolated from soil (HI2424) and from a CF patient (AU1054). This mapping step was performed such that only reads that map to a unique location within the reference were kept, meaning that repeated regions within the reference would not be covered by this step. While most reads were able to be mapped to all three reference genomes, both the average genome coverage (amount of the genome covered) and the fold coverage (average number of reads that align to any given position in the genome) were substantially higher when aligned to the J2315 strain (~96% genome coverage at ~300X fold coverage), compared with the two PHDC lineage strains (~80% genome coverage at ~210X fold coverage), as would be expected for ET-12 lineage strains (Table 4.3). Subsequent comparisons were thus performed only with reference strain J2315.

Table 4.3. Genome and fold coverage results when mapping Illumina reads of four newly sequenced clinical isolates against three reference genomes (AU1054, HI2424, and J2315), without allowing mapping to repeat regions in the reference strain.

AU16956		Total untrimmed reads	21,841,221
Reference	**AU1054**	**HI2424**	**J2315**
Aligned reads	12,494,813	12,917,926	18,249,087
chr1 coverage	86.6 % (213.3x)	88.6 % (219.3x)	94.7 % (283.9x)
chr2 coverage	80.5 % (201.1x)	80.9 % (203.7x)	98.9 % (324.0x)
chr3 coverage	80.7 % (186.9x)	69.3 % (153.9x)	97.6 % (321.0x)
plasmid coverage	No plasmid	34.4 % (24.3x)	96.9 % (177.4x)
overall coverage	83.3 % (204.3x)	81.8 % (200.1x)	96.7 % (302.7x)
AU16958		**Total untrimmed reads**	**22,839,719**
Reference	**AU1054**	**HI2424**	**J2315**
Aligned reads	13,675,442	14,015,382	20,314,513
chr1 coverage	83.8 % (231.1x)	88.5 % (249.6x)	94.6 % (318.5x)
chr2 coverage	80.4 % (217.7x)	80.8 % (221.4x)	98.5 % (347.1x)
chr3 coverage	80.4 % (211.8x)	60.6 % (137.7x)	97.4 % (353.2x)
plasmid coverage	No plasmid	20.5 % (37.7x)	96.8 % (280.6x)
overall coverage	82.0 % (222.8x)	80.3 % (218.7x)	96.5 % (333.3x)
HI4277		**Total untrimmed reads**	**24,170,663**
Reference	**AU1054**	**HI2424**	**J2315**
Aligned reads	14,059,449	14,506,045	20,719,225
chr1 coverage	81.6 % (248.9x)	88.0 % (269.4x)	94.1 % (340.1x)
chr2 coverage	80.1 % (210.0x)	80.5 % (214.4x)	98.7 % (336.6x)
chr3 coverage	79.9 % (213.9x)	55.0 % (130.3x)	97.7 % (331.2x)
plasmid coverage	No plasmid	19.4 % (32.6x)	96.9 % (244.5x)
overall coverage	80.8 % (228.2x)	79.1 % (223.8x)	96.3 % (336.6x)
HI4278		**Total untrimmed reads**	**23,729,743**
Reference	**AU1054**	**HI2424**	**J2315**
Aligned reads	13,802,908	14,267,350	20,561,601
chr1 coverage	81.9 % (239.5x)	88.5 % (260.2x)	94.3 % (327.8x)
chr2 coverage	78.7 % (211.8x)	79.2 % (216.1x)	97.3 % (341.4x)
chr3 coverage	80.1 % (214.4x)	54.5 % (131.9x)	97.7 % (343.4x)
plasmid coverage	No plasmid	20.4 % (37.4x)	96.9 % (279.6x)
overall coverage	80.4 % (224.8x)	78.7 % (220.7x)	95.9 % (334.3x)

Alignments of all reads to the J2315 genome was performed chromosome by chromosome (Table 4.4), and allowing mapping reads to repeats in order to ascertain genuine gaps within the newly sequenced genomes. Two methods were employed to identify possible gaps with respect to the reference genomes (ie. deemed to be absent from the sampled genome): 1) based on regions without a single read aligned; 2) based on regions that are covered less than 0.1 fold the average coverage. These "gaps" were totaled and compared to read-mapping without allowing alignments to repeat regions, giving an idea of the level of repetitiveness within the J2315 reference genome (Table 4.4).

While these numbers provide averages and a general notion of genome coverage, a more detailed picture of the sequence alignments is required to appreciate the differences in coverage levels along the genome (see Figure 4.2 for HI4277 and HI4278 data aligned to J2315 chromosome 1 as an example). It is clear from such alignments that while some regions are absent from HI4277, other regions of the J2315 reference appear to be overrepresented in the newly sequenced genome. When allowing reads to align to repeat regions, many spikes of overrepresented sequence arise, since reads can be aligned to multiple places (Figure 4.2A), whereas when this feature is disabled, repeat regions now appear as missing regions within the newly sequenced strains (Figure 4.2B,C). These regions have been confirmed to belong to classes of IS elements, and to a repeated family of retrotransposable group II intron elements. In addition, the 57kb duplication in chromosome 1 is confirmed and indicated as large repeats (black bars) at the bottom of Figure 4.2, are seen as overrepresented sequence when allowing repeat-mapping (Figure 4.2A), but appear as missing regions when ignoring reads that map to more than one location (Figure 4.2B,C).

Table 4.4. Read-mapping results of 4 novel strains versus *B. cenocepacia* J2315*

A) J2315 Coverage	AU16956	AU16958	HI4277	HI4278
Total aligned reads	15379193	16890096	17006158	16895618
Chr1 coverage (no repeat-mapping)	94.6 % (181.9x)	94.5 % (201.6x)	93.8 % (221.5x)	94.2 % (216.7x)
Chr1 coverage (multiple hits)	100.0 % (244.0x)	99.9 % (296.3x)	99.3 % (335.4x)	99.5 % (321.5x)
Chr2 coverage (no repeat-mapping)	98.8 % (203.3x)	98.4 % (214.4x)	98.6 % (215.2x)	97.0 % (221.8x)
Chr2 coverage (including repeats)	100.0 % (231.1x)	99.6 % (248.2x)	99.8 % (253.6x)	98.3 % (258.5x)
Chr3 coverage (no repeat-mapping)	97.4 % (205.6x)	97.1 % (222.6x)	97.4 % (215.8x)	97.5 % (227.0x)
Chr3 coverage (including repeats)	100.0 % (259.2x)	100.0 % (294.3x)	99.9 % (307.4x)	100.0 % (309.5x)
Plasmid coverage (no repeat-mapping)	96.8 % (127.7x)	96.9 % (200.2x)	96.9 % (178.0x)	96.9 % (204.6x)
Plasmid coverage (including repeats)	100.0 % (181.1x)	100.0 % (330.8x)	100.0 % (335.3x)	100.0 % (331.5x)
Overall coverage (no repeat-mapping)	96.6 % (192.4x)	96.3 % (209.0x)	96.2 % (217.9x)	95.7 % (219.7x)
Overall coverage (including repeats)	100.0 % (239.8x)	99.8 % (277.3x)	99.5 % (299.7x)	99.1 % (295.2x)
B) # Gaps (bp):	**AU16956**	**AU16958**	**HI4277**	**HI4278**
Chr1 Zero coverage	47 (861)	85 (4,588)	108 (28,905)	52 (17,805)
Chr1 <0.1 average	1599 (47,194)	2245 (80,237)	1596 (82,480)	1421 (63,225)
Chr2 Zero coverage	12 (164)	56 (11,965)	36 (6,757)	156 (53,256)
Chr2 <0.1 average	1018 (23,991)	1342 (50,512)	932 (29,922)	773 (81,605)
Chr3 Zero coverage	8 (108)	13 (415)	11(839)	4 (13)
Chr3 <0.1 average	308 (7,664)	403 (12,390)	303(8,827)	244 (6,893)
Pd Zero coverage	0 (0)	1 (39)	1 (20)	1 (19)
Pd <0.1 average	20 (579)	37 (921)	26 (700)	29 (634)

* Chr: Chromosome; Pd: Plasmid; A) No repeat mapping: only keep reads that map once; multiple hits: count all mappable locations for reads; B) Allowing multiple hits; Zero coverage: regions without any mapped reads; <0.1 average: all bases mapped by <10% of the average fold coverage

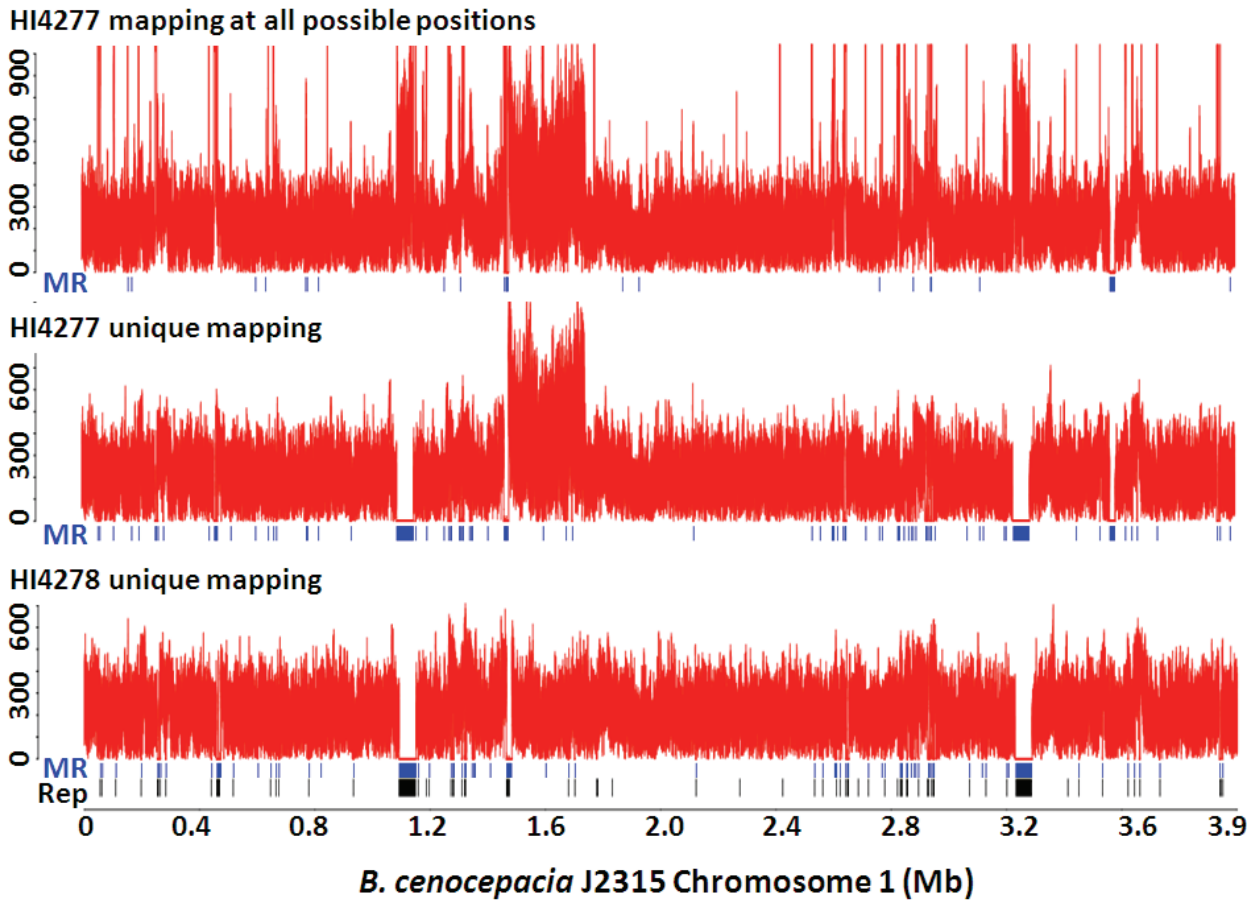


Figure 4.2. Genome and fold coverage read-based analysis. Examples of genome coverage where reads are mapped to the J2315 reference Chromosome 1. The fold coverage is represented by height of the peaks, large gaps are visible as zero coverage regions and further highlighted below each coverage plot by blue bars to represent missing regions >250bp (MR), and all repeats found in the J2315 genome are displayed at the bottom in black bars (Rep). A) HI4277 with reads mapped to one or more than one location (ie. if there are regions within the reference that are repeats within the HI4277, all these reads will map to the location(s) in the reference, showing up as deeper coverage regions); B) HI4277 with only reads that map uniquely to one location in the reference (ie. repeats will not be covered); C) HI4278 with only reads that map uniquely to one location.

In order to determine whether the missing or overrepresented regions are the same or different among the four strains, the precise coordinates of these presumed gaps were compared (MR in Figure 4.2; and Figure 4.3). One region overrepresented only in HI4277 appears to be a large duplication of ~200kb (Figure 4.2) that encompasses a large number of genes of varied function. Strain AU16956 has fewer gaps than the other strains (Figure 4.3), which explains the high degree of reference genome coverage observed (Table 4.4). Despite this fact, many of the missing regions with respect to the reference are shared among two or more of the sequenced strains, which supports the notion of a common evolutionary ancestor and history.

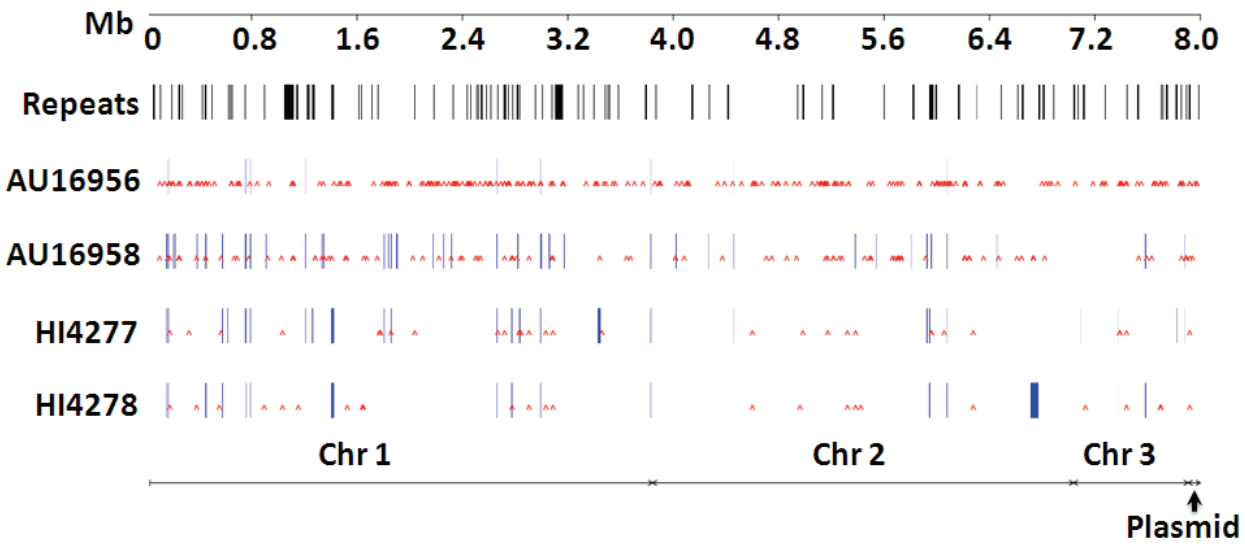


Figure 4.3. Regions of difference from the four sequenced strains compared with J2315. Gaps (blue bars) and SNPs (red carets) are indicated for each of the sequenced strains as well as their respective genomic locations (Mb) with respect to reference strain J2315. Only gaps of sufficiently large size (>250bp) to be visualized are illustrated.

A number of differences can be seen among the newly sequenced strains however, including those that would explain differences in J2315 genome coverage and total size of gaps. For example, HI4277 has a slightly smaller reference chromosome 1 coverage than the other strains (Table 4.4) and has a unique 19.4 kb gap near the 3.5 Mbp marker (Figure 4.3). Similarly, HI4278 displays the largest difference, >1% less coverage of chromosome 2 than its relatives (Table 4.4) which can readily be observed as its largest gap (a 63.6kb deletion), unique compared with the other strains (Figure 4.3). Both these strain-specific deletions, along with a few other deletions carry a number of proteins of diverse function (Figure 4.4) whose best blast hits are to members of the *Burkholderia* genus. Transcriptional regulators, which may have a number of downstream effects, were the largest group of affected functions (COG K) consistently in all strains, followed to a lesser extent by signal transduction (COG T) and cell motility (COG N). The number of genes affected by these missing regions and those shared between these four newly sequenced strains is displayed in Figure 4.5. While a large fraction of the affected genes are shared among more than one strain, and indeed 12 genes are commonly affected in all strains (Table 4.5), there are still a number of uniquely affected genes in each strain, as reflected in the Venn diagram (Figure 4.5).

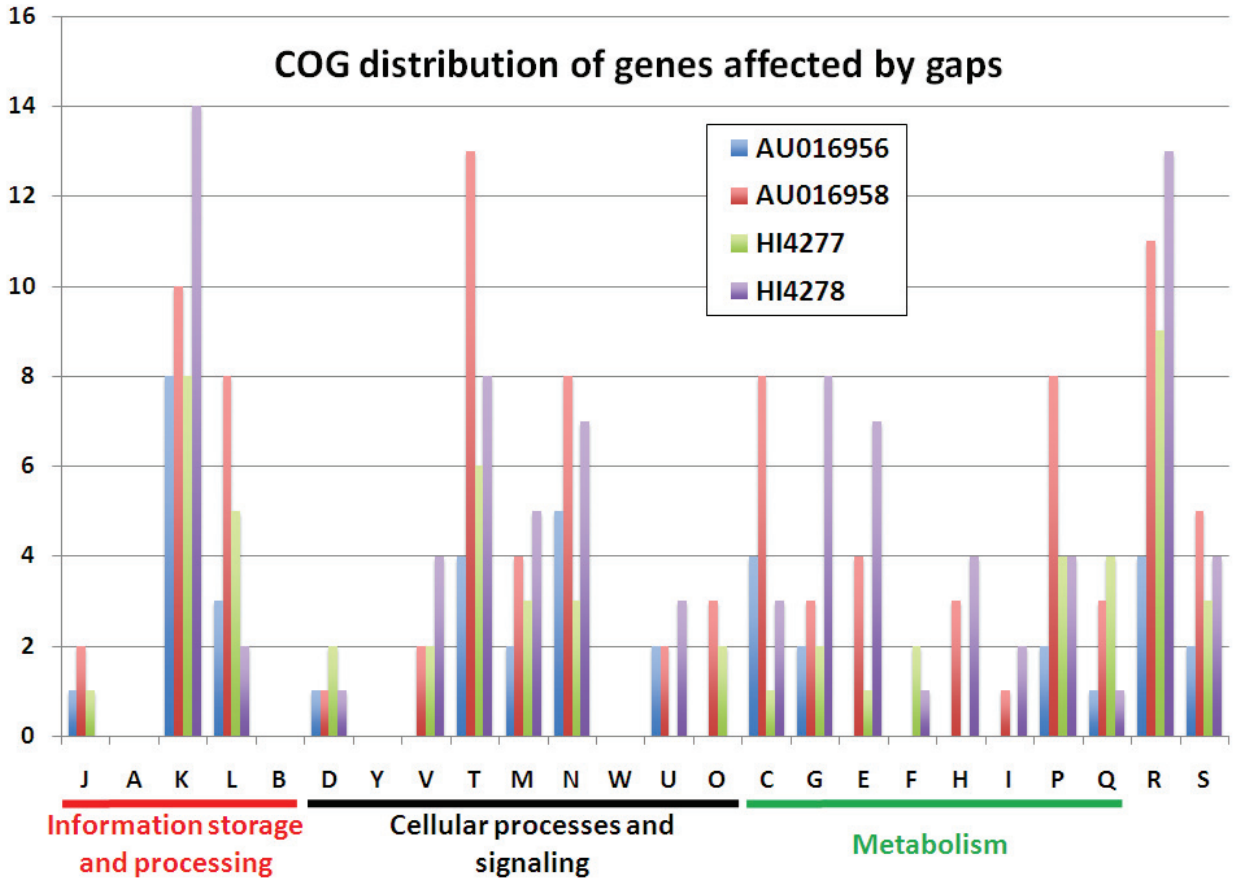


Figure 4.4. Distribution of genes affected by gaps in terms of functional classification according to the Clusters of Orthologous Groups (COG) of proteins. The number of J2315 genes that are missing from the newly sequenced genomes (as determined by read-mapping results, excluding repeat regions, and including regions with no coverage) are displayed by COG functional categories. Categories are as follows, J: Translation, ribosomal structure and biogenesis; A: RNA processing and modification; K: Transcription; L: Replication, recombination and repair; B: Chromatin structure and dynamics; D: Cell cycle control, cell division, chromosome partitioning; Y: Nuclear structure; V: Defense mechanisms; T: Signal transduction mechanisms; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; Z: Cytoskeleton; W: Extracellular structures; U: Intracellular

trafficking, secretion, and vesicular transport; O: Posttranslational modification, protein turnover, chaperones; C: Energy production and conversion; G: Carbohydrate transport and metabolism; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R: General function prediction only; S: Unknown.

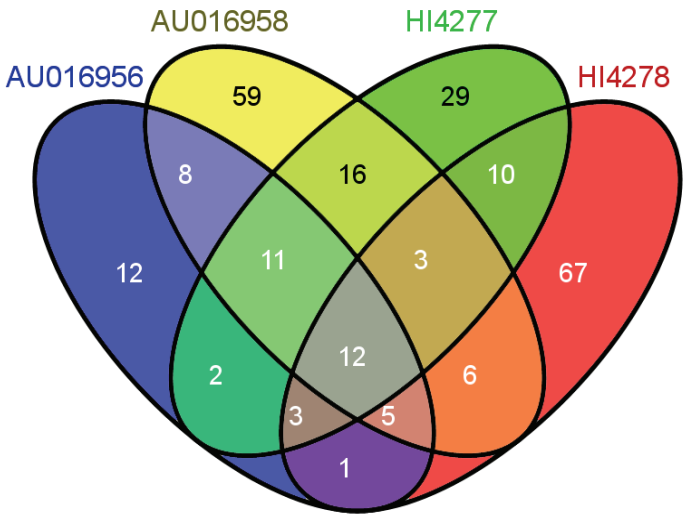


Figure 4.5. Venn diagram of genes affected by missing regions within the four sequenced ET-12 strains. While each strain has a number of gaps that uniquely affect a number of genes within the genome (ie. are different than the reference J2315), many of the genes affected by gaps are shared with other strains (e.g. over 75% of the genes affected by a gap in AU16956 are also affected in at least one of the other three strains). Twelve genes are affected in all four strains (see Table 4.5).

Table 4.5. List of genes in all four strains that are not entirely covered by read mapping.

J2315 Ortholog	Product description
BCAL0129	chemotaxis two-component sensor kinase CheA
BCAL0131	methyl-accepting chemotaxis protein
BCAL0196	Major Facilitator Superfamily protein
BCAL1389	cellulose synthase regulator protein
BCAL1893	family M23 peptidase
BCAL1899	DNA polymerase III subunits gamma and tau
BCAL2052	binding-protein-dependent transport system inner membrane protein
BCAL2397	putative lipoprotein
BCAM0886	LysR family regulatory protein
BCAM2027	hypothetical protein
BCAM2050	type III secretion system protein
BCAM2165	putative beta-lactamase

While the molecular mechanism underlying some deletions is unclear, several regions adjacent to one or more insertion sequence (IS) elements are presumed to have undergone deletion as a function of homologous recombination between parallel flanking elements, as has been observed in other organisms (5). Although IS elements have been implicated in genomic rearrangements in *B. cenocepacia* (12), this is the first evidence presented for genomic deletions. Further, a number of other regions that were not present in one or more of the four sequenced ET-12 strains appear to be of foreign origin, encode integrases or other phage-related proteins, along with many hypothetical proteins, and are likely transient mobile elements.

Single nucleotide polymorphisms identified among newly sequenced strains

While the majority of the differences between J2315 and the more recent ET-12 isolates lie within these gaps, over 99% of the reference genome is accounted for by mapped reads. We probed this information-rich portion of the genome by examining smaller differences that would not necessarily present themselves as uncovered or missing regions in read-based alignments, but where SNPs or small (<6bp) insertion/deletions (indels) could be predicted with high confidence. With the majority of bases covered from 100-400 fold, all positions that disagree with the reference genome were tallied. In addition, because we had already observed that genetic changes could even occur during the growth of *Burkholderia* for sample preparation (6), we also examined positions where >20% of the reads disagreed with the majority at that position.

A total of 25-236 SNPs were identified in each of the newly sequenced ET-12 genomes (Table 4.6, Figure 4.3). This range in number of mutations contrasts sharply with the 100-233 small indels observed in each of the new strains. It is interesting to note that there is a positive correlation with respect to the number of SNP and indel mutations, and that the strains isolated more than ten years ago have a larger number of both mutation types than do the more recently isolated strains (Table 4.6). The distribution of SNPs (Figure 4.3, Figure 4.6) and indels (not shown) in the genome appears random with respect to chromosome location and does not correlate with sequencing depth of coverage (Figure 4.6). However, examination of genes affected within each strain reveals that a small number of identical mutations are shared between strains, the pattern of which

corroborates other evidence of a common evolutionary ancestry for the two more recent isolates and the two older isolates (Figure 4.7).

Table 4.6. Number of SNPs and small indels found in each newly sequenced ET-12 strain of *B. cenocepacia*.

Replicon and feature	AU16956	AU16958	HI4277	HI4278
Chr1 indels **SNPs (S/NS/I)***	112 23/65/28	87 3/35/9	67 3/11/3	54 1/10/2
Chr2 indels **SNPs (S/NS/I)***	101 25/55/8	56 6/28/5	36 1/5/3	31 0/4/2
Chr3 indels **SNPs (S/NS/I)***	17 9/14/4	19 3/4/1	9 1/0/2	14 2/3/0
Plasmid indels **SNPs (S/NS/I)***	3 2/1/2	0 0/1/1	1 0/0/1	1 0/0/1
Total indels in Genome	233	162	113	100
Total SNPs in Genome	236	96	30	25

*S: Synonymous SNPs; NS: non-synonymous SNPs; I: intergenic SNPs

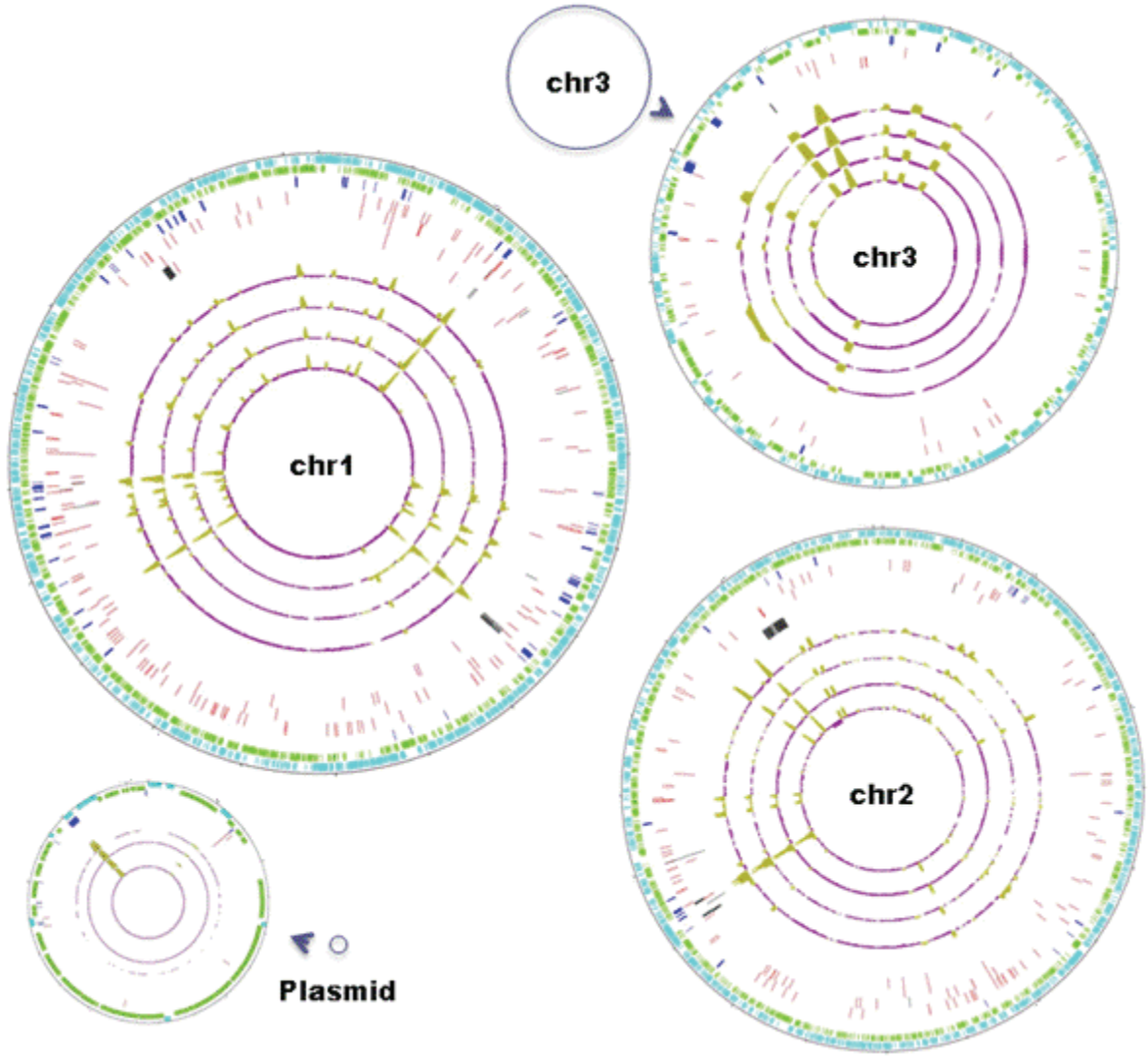


Figure 4.6. Read-mapping and SNPs found in four clinical isolates of *B. cenocepacia*. The circular representations of the J2315 genome with chromosome 3 and the plasmid enlarged (their respective sizes are outlined in empty blue circles). The rings (outer to inner) represent: all genes predicted in J2315 (rings 1 and 2); repetitive sequences in blue (ring 3); SNPs (red) and gaps (black) found in the four strains (AU016958, AU016956, HI4278, and HI4277 in rings 4-8, respectively); and read-mapping results with respect to average coverage, and allowing multiple hits per read (green is overabundant, and purple is underrepresented, same order as SNPs, rings 9-12).

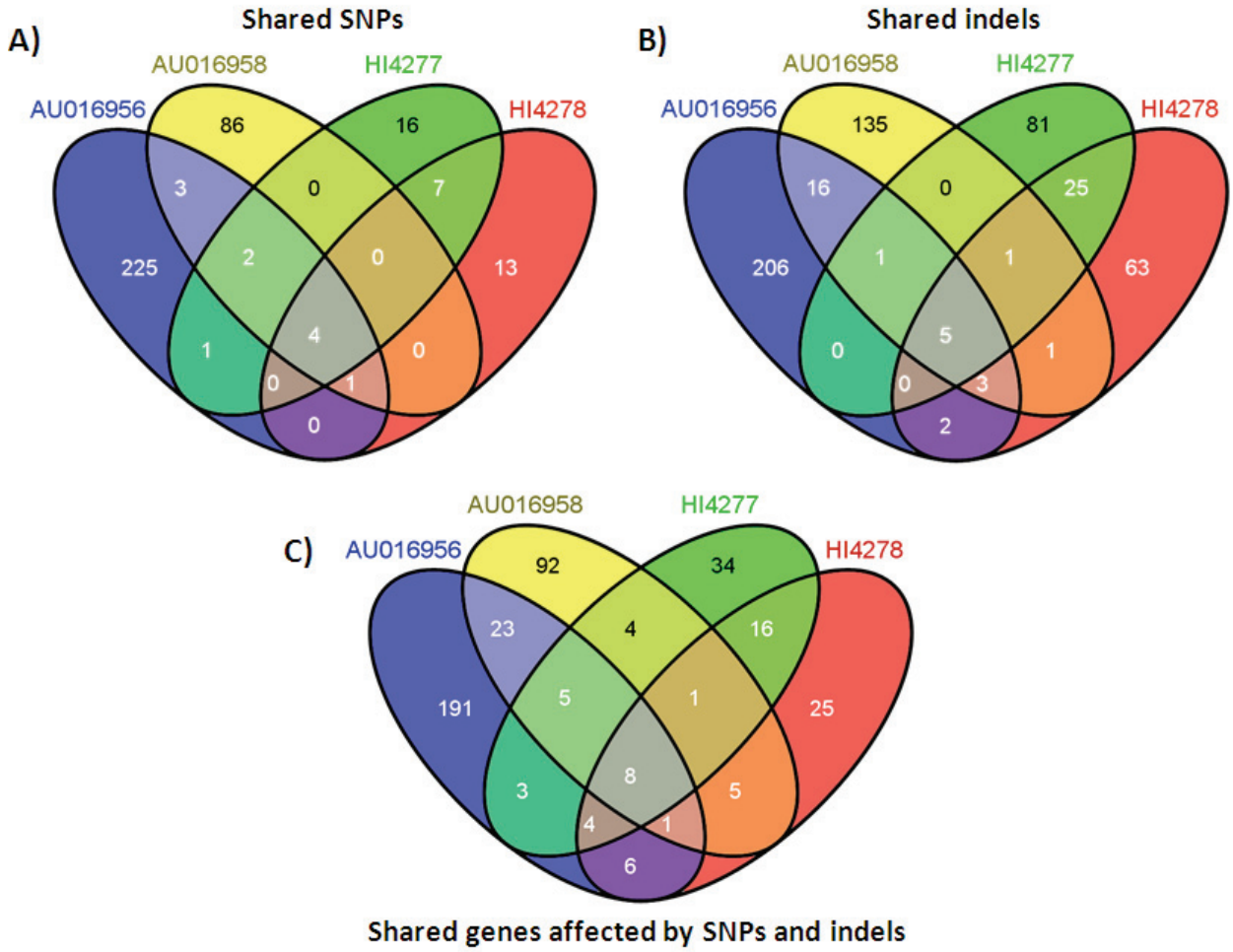


Figure 4.7 Shared SNPs, shared indels, and shared genes that are perturbed by these mutations, within four ET-12 strains. Counts of those genes that have unique and shared SNPs (A) or indels (B) are illustrated in Venn diagrams (ie. different strains have the same mutation – SNP or indel – with respect to the J2315 reference). Also presented are genes that are affected by either non-synonymous SNPs or indels that affect the coding sequence (C). While the majority of both mutation types are unique to each strain, a small subpopulation of the mutations are shared between the two presumed lineages of ET-12 (see text), and an even smaller number are shared among all four strains.

While the majority of SNPs and indels were found to be strain-specific, highlighting the rapid evolution that may take place within a host, a number of identical mutations were found between two or more strains (Figure 4.7, and SNPs in Table 4.7), and support the presumed evolutionary history of the four strains: AU16956 and AU16958, two strains isolated from the same hospital epidemic two years apart (1999 and 2001) appear to share a common lineage; as do HI4277 with HI4278, two strains isolated in 2008 from the same hospital. When looking at genes affected by non-synonymous SNPs and indels, it appears that a number of these mutations occurred in the same gene within two or more of the newly sequenced strains, but that the mutations were at different locations within the gene (Figure 4.7). With the assumption that these mutations may similarly alter or abolish protein function, this perhaps suggests convergent evolutionary forces acting on specific functions. The other reason why the numbers of affected genes per strain are smaller than the sum of the SNP and indel mutations is that a number of the indels and SNPs occur in intergenic regions, including within genes that have already been pseudogenized via other mutations also present in J2315 (e.g. interruption via IS elements), and are not recognized as protein-encoding genes.

Genes with SNPs, as analyzed by COG functional classification, were scrutinized with respect to the type of SNPs they carried, however all functional categories were found to have accrued more non-synonymous mutations than ones that did not alter coding sequence. The functional breakdown of all genes altered by either indels or non-synonymous SNPs (Figure 4.8) illustrates no predominant functional group affected by these mutations. It is interesting to note the contrast between the number of genes affected by indels and SNPs, and the number of genes affected by gaps (Table 4.4, Figure

4.4) in each strain. For example, AU16956 has the fewest gap regions, in number, in total bp, and in genes affected by gaps, yet has the largest number of genes affected by non-synonymous SNPs and indels.

Table 4.7. Identical SNPs in newly sequenced strains compared with J2315.

J2315 Ortholog	Product description	SNP Type	Strains that share SNP
BCAL0155	putative cation efflux protein	NS	All four strains
BCAL2656	cobyric acid synthase	NS	All four strains
BCAL2830	two-component regulatory system, sensor kinase protein	S	All four strains
BCAL0356	putative quinone oxidoreductase	NS	AU16956, AU16958, HI4278
BCAL2475	hypothetical protein	NS	AU16956, AU16958, HI4277
BCAM1223	hypothetical protein	S	AU16956, AU16958, HI4277
BCAS0609	putative electron transfer flavoprotein-ubiquinone oxidoreductase	S	AU16956, HI4278
BCAL0520	putative flagellar hook-length control protein FliK	NS	AU16958, HI4277
BCAM0691	hypothetical protein	NS	HI4277, HI4278
BCAM2193	putative 3-hydroxyisobutyrate dehydrogenase	NS	HI4277, HI4278
BCAS0347	putative binding-protein-dependent transport system protein	NS	HI4277, HI4278
BCAL0954	recombination regulator RecX	S	HI4277, HI4278
BCAL2784	hypothetical protein	S	HI4277, HI4278
BCAM1397a	putative squalene/phytoene synthase	S	HI4277, HI4278
BCAL2188	putative single-stranded-DNA-specific exonuclease	NS	AU16956, AU16958
BCAS0747	hypothetical protein	NS	AU16956, AU16958
BCAL1391	putative cellulose biosynthesis protein	S	AU16956, AU16958
BCAM1210	glutamine ABC transporter ATP-binding protein	S	AU16956, AU16958
BCAM2143	cable pilus associated adhesin protein	S	AU16956, AU16958

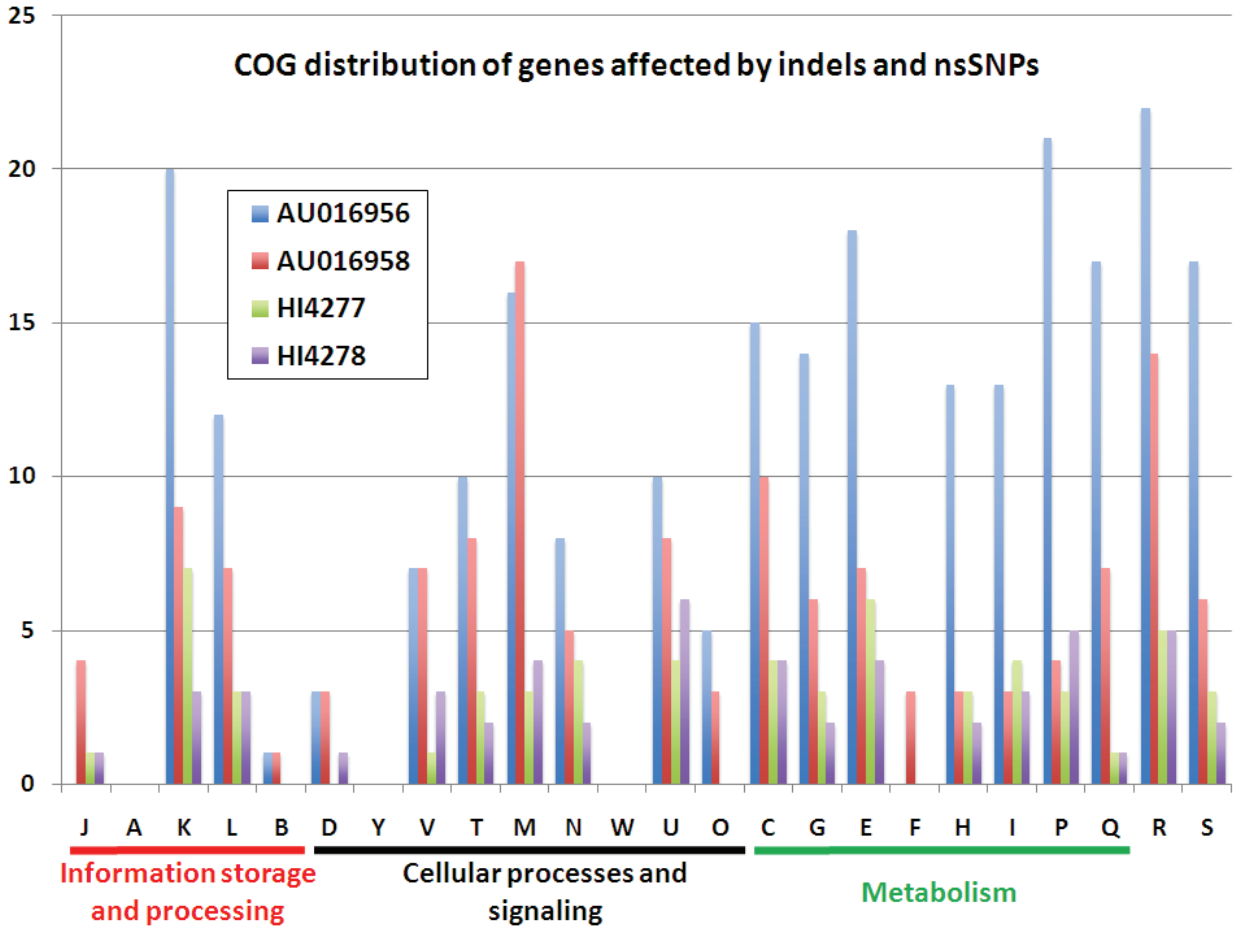


Figure 4.8. Distribution of genes altered by non-synonymous SNPs or indels, in terms of functional classification according to the Clusters of Orthologous Groups (COG) of proteins. The genes with non-synonymous SNPs or indels when compared with reference J2315 (as determined by read-mapping results) are displayed by COG functional categories for each sequenced strain. Categories are as designated in the figure legend of Figure 4.4.

De novo assembly reveals novel ET-12 genomic DNA

While mapping reads against reference genomes does allow for highly refined analyses of polymorphisms and missing regions in newly sequenced strains, it is harder to gain information on genomic novelty found in the short read sequence information. Toward this end, we used a number of bioinformatics approaches to assemble the data for comparison with the reference J2315. Although partially expected given a large genome with many repetitive regions, the assemblies of all four datasets produced thousands of contigs that could be adequately aligned to the reference genome, including the plasmid (Table 4.8, Figure 4.9).

Table 4.8. Draft assembly and novel ET-12 regions not present in strain J2315.

	AU16956	AU16958	HI4277	HI4278
Assembled contigs	3,251	3,703	2,540	2,656
Total bases in contigs	8,287,871	7,998,367	7,989,566	7,895,668
Largest contig	52,547	52,520	92,869	75,214
Contig N50	7,458	6,219	9,439	8,873
Total size in contigs >5kb	5,602,838	4,744,437	6,096,700	5,835,243
# contigs >5kb	536	480	541	528
Novel regions >500bp	5	1	1	1
Novel DNA in bp	11,966	1,629	1,629	1,629

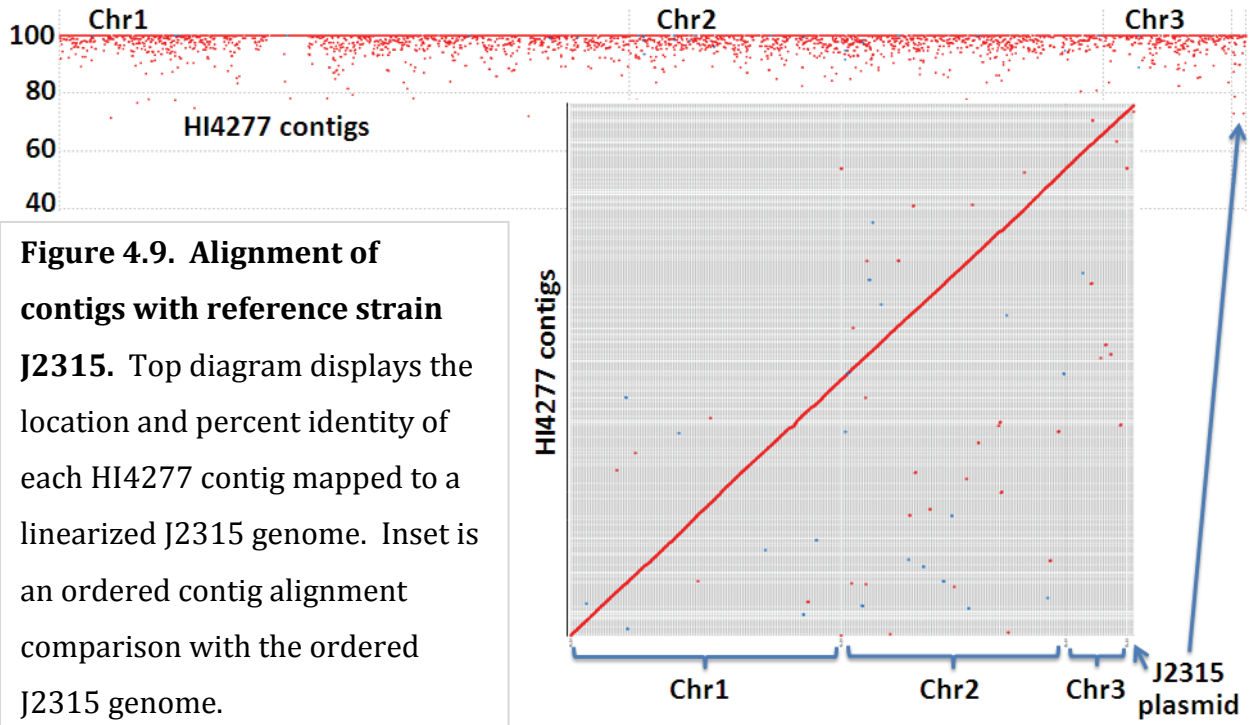


Figure 4.9. Alignment of contigs with reference strain J2315. Top diagram displays the location and percent identity of each HI4277 contig mapped to a linearized J2315 genome. Inset is an ordered contig alignment comparison with the ordered J2315 genome.

The assembly metrics for all four new strains were similar, with contigs between 6 and 10kb representing 50% of the assembly and presumably >50% of the ET-12 genomes (Table 4.8). The contigs were also aligned to the reference J2315 genome and these results largely confirmed the SNP and gap observations from read-mapping experiments. Some differences were observed and verified manually; most discrepancies arose from the fact that contigs may span regions with high diversification, resulting in SNPs that appear as gaps with read-mapping. Other differences included repeat regions which appeared as gaps in the assembly.

Contigs that mapped well to the J2315 genome were examined for regions >500bp that did not match the reference. The few regions identified were further examined for similarity to known sequences in Genbank using blastN and blastX (Table 4.9). The majority of novel ET-12 DNA was found in the AU16956 strain, and most novel ET-12 DNA

had hits to other *Burkholderia* spp., including lineages of *B. cenocepacia*, or in other species of the *Burkholderia cepacia* complex. One region was found in all four strains and encodes a phage tail protein most similar to one found in *B. pseudomallei*, and a hypothetical protein most similar to *B. multivorans*. One region in AU16956 has only been reported outside the *Burkholderia*, a 647bp region most similar to a conserved hypothetical protein in *Verminephrobacter* (Table 4.9). Given that these regions were discovered in contigs which have partial matches to the J2315 reference genome, we presume these are indeed part of the genomes of the four newly sequenced strains. In fact, two of the novel ET-12 regions found in AU16956 may be part of a Chromosome 3 island present in non-ET-12 strains of *B. cenocepacia* (Figure 4.10). Interestingly, this region is found beside the BcepMu prophage which itself is a known island in ET-12 strains (39), thus AU16956 may represent an intermediate strain which harbors both islands (Figure 4.10).

Table 4.9. Novel ET-12 lineage DNA and encoded products.

Strain	Size (bp)	Closest blast	Product/Description
AU16956	4173	betaproteobacteria	*trbLFGI* conjugal transfer genes
AU16956	2972	*B. cenocepacia* strains	benzaldehyde dehydrogenase; Alcohol dehydrogenase
AU16956	2557	*B. cenocepacia* strains	Amidohydrolase 3; transcriptional regulator
AU16956	647	*Verminephrobacter*	Conserved hypothetical protein
AU16956	1,617	*B. pseudomallei*; *B. multivorans*	Phage tail protein; hypothetical
AU16958	1,629	*B. pseudomallei*; *B. multivorans*	Phage tail protein; hypothetical
HI4277	1,629	*B. pseudomallei*; *B. multivorans*	Phage tail protein; hypothetical
HI4278	1,629	*B. pseudomallei*; *B. multivorans*	Phage tail protein; hypothetical

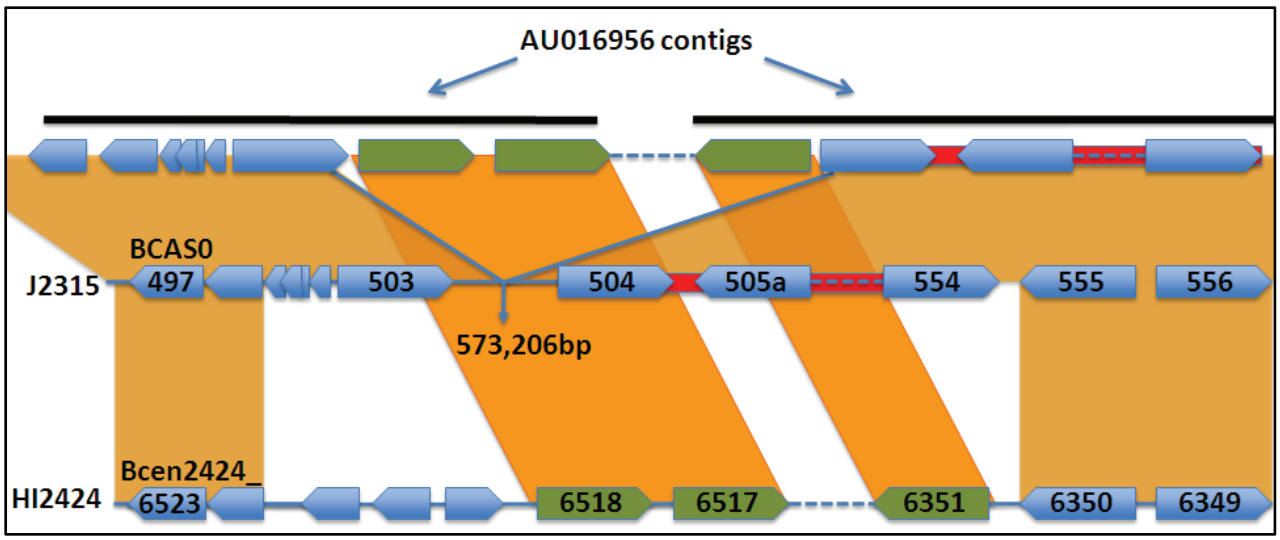


Figure 4.10. Comparison of two AU16956 contigs with the genomes of *B. cenocepacia* lineage ET-12 strain J2315 and PHDC lineage strain HI2424. Two contigs (black horizontal lines) from the assembly of AU16956 that aligned well to the J2315 reference genome also contained DNA not found in the reference. The 3.0kb and 2.6kb regions (see Table 4.9) were found to have high similarity to regions found in many other, non-ET12 lineage *B. cenocepacia* (HI2424 is shown as an example). Wide arrows represent genes, green colored arrows represent good blast matches to the contig ends, hatched lines indicate large regions of DNA between genes (and unknown sequence length between the contigs of AU16956). Horizontal red bars represent the BcepMu prophage, orange and brown color between the genomes represent regions of homology. Gene numbers for select genes are indicated for J2315 (BCAS0 gene labels) and HI2424 (Bcen2424_ gene labels) as reference.

Discussion

As the array of competitive and high throughput sequencing platforms continues to diversify, the number and variety of bacterial genome projects continues to grow. In addition, as the use of these new machines becomes more common in institutions other than the world's few dominating sequencing centers, including in university settings, the use of sequencing for whole genome sequencing and resulting genome quality has become a concern (7). With the Illumina and SOLiD platforms able to sequence multiple bacterial genomes at deep coverage in a single run, their use in bacterial genomics is now widespread. The short reads and remaining quality issues associated with these data have prevented these platforms from completely overtaking previous sequencing approaches for whole genome sequencing however, and have partially limited their use to re-sequencing applications (20), including sequencing recently evolved strains (e.g. (4, 18)) or for transcriptomic mapping/RNAseq (e.g. (45)) efforts. Here, we applied Illumina sequencing to determine the genome composition of four clinical isolates of the pathogenic *B. cenocepacia* ET-12 lineage, often associated with epidemic infections in Canada and Europe. Strains HI4277 and HI4278 are representative strains from the 2008 Toronto outbreak (isolated in September 2008) and are of particular interest since the respective patients had very different outcomes (one died quickly, 55 days after the first positive culture, while the other was still alive ~1.5 years after positive culture). It is unclear whether the clinical outcome was primarily influenced by the host or pathogen, but is a question that could potentially be answered through genomic analysis. The two other strains are from a prior outbreak in the same location and were isolated in 1999 (strain AU016956) and 2001 (strain AU016958). The potential to examine evolutionary change in

circulating strains within a clinical setting over a decade and the ability to compare with another ET-12 clone prompted the draft Illumina sequencing of these strains and subsequent analyses presented here.

Read-mapping based comparisons to ET-12 J2315 showed that chromosomes 2 and 3 were consistently better covered than chromosome 1, which is in stark contrast to identical comparisons against reference PHDC genomes HI2424 and AU1054 (Table 4.3). Although it is known that AU1054 harbors a translocation between chromosomes 1 and 3 (17), this is not sufficient to explain the large difference in genome coverage statistics. Instead, this observation likely results from the fact that the various chromosomes appear to be under different selective pressure and evolve at different rates (6, 10). The two PHDC lineage strains have sufficiently diverged that reads cannot be evenly aligned among the chromosomes using strict alignment tools, specifically highlighting the rapid evolution of chromosomes 2 and 3, and underscoring the great diversity of the *Burkholderia* genus.

In refined comparisons to the J2315 genome, a large 57kb duplicated region along with a number of repeated insertion sequence (IS) elements have been confirmed present within the four new strains, similar to J2315 (21), although the precise copy number of the smaller repeats can only be approximated. A large duplication was found, but only in the HI4277 strain (Figure 4.2). In terms of deletion events, based on read-mapping it appears that the majority of gene flux occurs in chromosomes 1 and 2, with large deletions (compared with J2315) occurring in these chromosomes, particularly within the more recently isolated strains (HI4277 and HI4278). These two strains may have acquired these additional large scale deletions throughout generations of replication and evolution within the cystic fibrosis community for years through patient to patient spread. We speculate

that such large deletion events occur in a neutral fashion within this species, which is well-adapted to its niche, and can accommodate the loss of certain functions that are no longer required for survival outside its host. Other regions missing from the newly sequenced strains yet present in the reference appear to be horizontally acquired material of phage origin, and may have been recently deleted from these strains or acquired by J2315. Only 12 genes were affected by missing regions in all four strains, and these include two chemotaxis related proteins, transport proteins, a type III secretion system protein, a peptidase, a transcriptional regulator and a beta lactamase. While the loss or alteration of chemotaxis functions may explain the lack of motility recently observed in some ET-12 strains such as HI4277 (38), it is unclear what effects the other mutations have on the phenotypes of these strains.

Despite these large differences, the majority of the genomes of the four sequenced strains are highly similar to the reference J2315 genome. Indeed, aside from the larger deletions described above, only 25-236 SNPs and 100-233 small indels were found in any of the new ET-12 strains, many of which amplify the effect of other mutations shown to have already pseudogenized genes (21). While it is unclear if any of the non-synonymous SNPs contribute to altered protein function, one of these affects the sequence of a cation efflux pump protein, which has been implicated in chlorhexidine resistance in *Klebsiella pneumonia* (13). This candidate mutation is of interest in terms of its effect on native high resistance of ET-12 isolates to chlorhexidine (35), given that *B. cepacia* complex (*Bc*c) strains have been implicated in nosocomial outbreaks, including some due to chlorohexidine contamination (34). Similarly, the insertion of a single nucleotide (resulting in a frameshift) in all newly sequenced strains, appears to have disrupted a penicillin-

binding protein precursor (BCAL2832 in J2315) that may affect their resistance to beta-lactam antibiotics.

Assemblies of the four new datasets helped confirm read-mapping results, and in addition allowed the discovery of a few novel sequences that have until now, not been reported in ET-12 lineage strains. None of these regions appear to have been acquired recently, as the vast majority of these regions are orthologous to regions present in other sequenced *B. cenocepacia* strains or *Bcc* strains. One interesting and notable observation is the fact that strain AU16956 was the only strain that carried more than one novel ET-12 region, including two regions only found in other *Burkholderia* spp. or other betaproteobacteria, and two other regions which are homologous to non-ET12 *B. cenocepacia* loci that are located adjacent to the insertion site of an ET-12-specific BcepMu island (Figure 4.9). These findings suggest that the J2315 reference and the three other newly sequenced ET-12 strains may have lost these regions. In addition, when reads are mapped to the J2315 genome, AU16956 appears to harbor fewer gaps compared with the other strains (Table 4.4, Figure 4.3). Furthermore, AU16956 has more observed SNP and small indel differences than the other strains (Figure 4.3, Table 4.6, Figure 4.7). Given these observations, it is tempting to speculate that this strain is of a different lineage than the other ET-12 strains, despite being isolated from the same hospital and only 2 years prior to AU16958. An alternative theory is that AU16956 was subject to strong selective pressure, possibly from treatment regimens, resulting in a high number of SNPs and small indels. This however, does not explain the lack of missing regions compared with AU16958, nor the novel ET-12 DNA whose only homologs lie within other *Burkholderia* or betaproteobacteria.

While the *Burkholderia* genus and *Bcc* organisms are recognized for their capacity to acquire and harbor DNA of foreign origin (6, 21, 40), it is also known that both free-living organisms and pathogens (within the context of a host) can evolve through the process of genome reduction (1, 31, 44). It appears that several *Burkholderia* spp. have evolved using this mechanism, including the genome of *B. rhizoxinica* that has undergone streamlining and is substantially smaller than its rhizosphere relatives (25). Another example is the highly reduced genome of *B. mallei*, a clonal lineage of *B. pseudomallei* and the causative agent of glanders, that is a highly adapted, obligate parasite of horses, mules and donkeys, with no other known reservoir (32). In *B. cenocepacia*, although similar in size to other *Bcc* members, a number of genetic loci do appear to be no longer functional in the J2315 genome (21). The lack of new genomic islands in the genomes of these four isolates, combined with the differential loss of both large and small, and seemingly random loci in the various genomes of more recent ET-12 isolates may be indicative of a genome streamlining strategy, as opposed to an evolutionary strategy based primarily on acquisition of novel genetic loci via horizontal gene transfer. Another interesting finding supports the notion of convergent evolution in these more recent ET-12 strains, since non-identical mutations including SNPs and indels have occurred in the same ortholog, possibly altering the same function in different strains. One interesting example, a member of a large family of outer membrane proteins that have been shown to play important roles in the pathogenesis of bacterial infections such as host cell attachment and invasion (19), is a putative haemagglutinin-related autotransporter protein (BCAM2418 in J2315) that has different mutations in all four strains.

The rapid high throughput sequencing of four new isolates of *B. cenocepacia* ET-12 and associated comparative analyses presented here further extend observations that this lineage is strongly associated with immunocompromised hosts (particularly those with CF), and suggest that the genome of the ET-12 lineage may be undergoing reductive evolution. This is supported by observations that J2315, isolated in 1989, carries a number of pseudogenes , that these pseudogenes are predicted to also be present in all four newer isolates, and that little to no new DNA material has been acquired by these strains despite missing a number of loci found in J2315. Most of the regions found in one or more of these newer strains that are not present in J2315 appear to be conserved in other *Burkholderia* and have likely undergone deletion in the J2315 lineage. Furthermore a number of new potential pseudogenes have arisen in each of the four strains, through both non-synonymous SNPs and small indels that in many cases have created frameshifts in genes. Further completion of these genomes will help corroborate and expand on the narratives touched upon above, and epidemiologic sequencing of a more diverse collection of ET-12 strains, isolated from different outbreaks in different geographic locations will shed light on whether the genome of this lineage of *B. cenocepacia* is indeed undergoing genome degradation while continuing to adapt to its niche, and if the ET-12 group include multiple distinct lineages that have been isolated from cystic fibrosis patients.

Perspective

The field of microbial genome sequencing is a frenzied one. Despite only being introduced in 2007, and being readily available since only 2008, the Illumina and other short read sequencers have become the dominant platform for genome sequencing applications. While the strides in sequencing throughput continues to outpace Moore's Law, the development of bioinformatics tools to process and analyze such data lag appallingly behind.

In this study, we used one of the first generation software upgrades with chemistry kits for obtaining longer, 76bp reads. As can be seen in Figure 4.1, the quality of such reads resulted in the equivalent of a fraction of the data being useable, with only ~75% of the length (~56-58bp average). In hindsight, an alternate strategy may have been better, as high sequence quality is of prime importance, and several rounds of data analysis were required to optimize the use of the reads. In order to adequately handle their quality, new tools were developed to process these vast amounts of data. While most detailed comparative genomics studies have been performed using finished genomes, with the current short read sequencers dominating the sequencing market, the ease with which raw sequence data can be generated, and the difficulties associated with genome finishing, it is clear that new software must be developed to handle short reads. There do exist several tools that can align short reads to a reference and even adequately parse the data to obtain SNPs and missing regions. Although this is at best mostly true for simple cases. The difficulty lies in determining the parameters and definitions in these processes. For example, despite an average fold coverage of above 200X, certain regions are still covered with only 5-20 reads. This can be due to non-specific alignments to otherwise "unique"

regions, or could alternatively mean that this region is sufficiently different from the reference that the alignment parameters fail to place the reads in this location adequately. It is thus critical to carefully weigh the desired straightforward outcome with the criteria used to perform the alignments. Although it is clear when there is a difference in a particular region, the distinction between SNPs, that could possibly not even affect coding sequence, versus indels which always affect coding sequence (if present within the context of a gene) is a significant one. In these cases, one would prefer to rely on assembled genomic sequence, which would bypass these difficulties if the regions of difference assemble well. As with most situations however, the combination of different analytical strategies is likely the most appropriate solution, if a touch overwhelming.

Assembly will also allow the capture of rearrangements, which are difficult for short read alignment programs, particularly if they occur within the context of repetitive regions which occurs frequently in bacteria. Unfortunately, assemblers are notoriously picky about the quality of reads that they assemble, and methods are still being developed to perform *de novo* assembly on such large datasets. In addition, short read assemblers that can handle millions of reads, generally produce fragmented assemblies with hundreds of contigs, in part due to the small size of the reads and overlaps which cannot resolve repetitive regions larger than the read lengths. Different tools are then required to attempt alignment of longer contigs to the reference genome. The tools and parameters used for such contig alignments are also important, and may affect the way larger contigs align. Unfortunately, most tools appropriate for read-based analyses are not appropriate for contig-based ones, resulting in less straightforward comparisons. Differences between read-based and contig-based alignments also exist since contigs may be able to overcome

difficulties in aligning reads to divergent regions, allowing alignment of longer contigs despite having internal regions that do not align well.

While this Illumina sequencing effort provided raw data that was far from ideal, the development of new methods for data processing coupled with manual inspection of genomic differences allowed a thorough examination of several new isolates of the highly pathogenic ET-12 lineage of *B. cenocepacia*, and have provided a framework for studying available and sequenced isolates of the same species. In addition, several major improvements in sequencing chemistry and engineering have occurred since the inception of this project, and together with new tools for data mining and analysis, it is clear that high-throughput sequencing will continue to hold tremendous potential for years to come, for studying bacterial evolution and population dynamics and for following the impact of genomic differences in the clinical outcomes of infected patients.

References

1. **Andersson, J. O., and S. G. Andersson.** 1999. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev **9:**664-71.

2. **Asghar, A. H., S. Shastri, E. Dave, I. Wowk, K. Agnoli, A. M. Cook, and M. S. Thomas.** 2010. The pobA gene of Burkholderia cenocepacia encodes a Group I Sfp-type phosphopantetheinyl transferase required for biosynthesis of the siderophores ornibactin and pyochelin. Microbiology.

3. **Baldwin, A., E. Mahenthiralingam, P. Drevinek, P. Vandamme, J. R. Govan, D. J. Waine, J. J. LiPuma, L. Chiarini, C. Dalmastri, D. A. Henry, D. P. Speert, D. Honeybourne, M. C. Maiden, and C. G. Dowson.** 2007. Environmental Burkholderia cepacia complex isolates in human infections. Emerg Infect Dis **13:**458-61.

4. **Barrick, J. E., D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim.** 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature **461:**1243-7.

5. **Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia.** 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A **101:**13826-31.

6. **Chain, P. S., V. J. Denef, K. T. Konstantinidis, L. M. Vergez, L. Agullo, V. L. Reyes, L. Hauser, M. Cordova, L. Gomez, M. Gonzalez, M. Land, V. Lao, F. Larimer, J. J. LiPuma, E. Mahenthiralingam, S. A. Malfatti, C. J. Marx, J. J. Parnell, A. Ramette, P. Richardson, M. Seeger, D. Smith, T. Spilker, W. J. Sul, T. V. Tsoi, L. E. Ulrich, I. B. Zhulin, and J. M. Tiedje.** 2006. Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A **103:**15280-7.

7. **Chain, P. S., D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter.** 2009. Genomics. Genome project standards in a new era of sequencing. Science **326:**236-7.

8. **Coenye, T., and P. Vandamme.** 2003. Diversity and significance of Burkholderia species occupying diverse ecological niches. Environ Microbiol **5:**719-29.

9. **Compant, S., J. Nowak, T. Coenye, C. Clement, and E. Ait Barka.** 2008. Diversity and occurrence of Burkholderia spp. in the natural environment. FEMS Microbiol Rev **32:**607-26.

10. **Cooper, V. S., S. H. Vohr, S. C. Wrocklage, and P. J. Hatcher.** 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput Biol **6:**e1000732.

11. **Dantas, G., M. O. Sommer, R. D. Oluwasegun, and G. M. Church.** 2008. Bacteria subsisting on antibiotics. Science **320:**100-3.

12. **Drevinek, P., A. Baldwin, L. Lindenburg, L. T. Joshi, A. Marchbank, S. Vosahlikova, C. G. Dowson, and E. Mahenthiralingam.** 2010. Oxidative stress of Burkholderia cenocepacia induces insertion sequence-mediated genomic rearrangements that interfere with macrorestriction-based genotyping. J Clin Microbiol **48:**34-40.

13. **Fang, C. T., H. C. Chen, Y. P. Chuang, S. C. Chang, and J. T. Wang.** 2002. Cloning of a cation efflux pump gene associated with chlorhexidine resistance in Klebsiella pneumoniae. Antimicrob Agents Chemother **46:**2024-8.

14. **Goldberg, J. B.** 2007. Polysaccharides of Burkholderia spp., p. 93-110. *In* T. Coenye and P. Vandamme (ed.), Burkholderia molecular microbiology and genomics. Horizon Bioscience, Wymondham, UK.

15. **Govan, J. R., P. H. Brown, J. Maddison, C. J. Doherty, J. W. Nelson, M. Dodd, A. P. Greening, and A. K. Webb.** 1993. Evidence for transmission of Pseudomonas cepacia by social contact in cystic fibrosis. Lancet **342:**15-9.

16. **Guglierame, P., M. R. Pasca, E. De Rossi, S. Buroni, P. Arrigo, G. Manina, and G. Riccardi.** 2006. Efflux pump genes of the resistance-nodulation-division family in Burkholderia cenocepacia genome. BMC Microbiol **6:**66.

17. **Guo, F. B., L. W. Ning, J. Huang, H. Lin, and H. X. Zhang.** 2010. Chromosome translocation and its consequence in the genome of Burkholderia cenocepacia AU-1054. Biochem Biophys Res Commun **403:**375-9.

18. **Harris, D. R., S. V. Pollock, E. A. Wood, R. J. Goiffon, A. J. Klingele, E. L. Cabot, W. Schackwitz, J. Martin, J. Eggington, T. J. Durfee, C. M. Middle, J. E. Norton, M. C. Popelars, H. Li, S. A. Klugman, L. L. Hamilton, L. B. Bane, L. A. Pennacchio, T. J.**

Albert, N. T. Perna, M. M. Cox, and J. R. Battista. 2009. Directed evolution of ionizing radiation resistance in Escherichia coli. J Bacteriol **191:**5240-52.

19. **Henderson, I. R., F. Navarro-Garcia, M. Desvaux, R. C. Fernandez, and D. Ala'Aldeen.** 2004. Type V protein secretion pathway: the autotransporter story. Microbiol Mol Biol Rev **68:**692-744.

20. **Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, J. I. Glasscock, M. Hickenbotham, W. Huang, V. J. Magrini, R. J. Richt, S. N. Sander, D. A. Stewart, M. Stromberg, E. F. Tsung, T. Wylie, T. Schedl, R. K. Wilson, and E. R. Mardis.** 2008. Whole-genome sequencing and variant discovery in C. elegans. Nat Methods **5:**183-8.

21. **Holden, M. T., H. M. Seth-Smith, L. C. Crossman, M. Sebaihia, S. D. Bentley, A. M. Cerdeno-Tarraga, N. R. Thomson, N. Bason, M. A. Quail, S. Sharp, I. Cherevach, C. Churcher, I. Goodhead, H. Hauser, N. Holroyd, K. Mungall, P. Scott, D. Walker, B. White, H. Rose, P. Iversen, D. Mil-Homens, E. P. Rocha, A. M. Fialho, A. Baldwin, C. Dowson, B. G. Barrell, J. R. Govan, P. Vandamme, C. A. Hart, E. Mahenthiralingam, and J. Parkhill.** 2009. The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol **191:**261-77.

22. **Isles, A., I. Maclusky, M. Corey, R. Gold, C. Prober, P. Fleming, and H. Levison.** 1984. Pseudomonas cepacia infection in cystic fibrosis: an emerging problem. J Pediatr **104:**206-10.

23. **Johnson, W. M., S. D. Tyler, and K. R. Rozee.** 1994. Linkage analysis of geographic and clinical clusters in Pseudomonas cepacia infections by multilocus enzyme electrophoresis and ribotyping. J Clin Microbiol **32:**924-30.

24. **Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg.** 2004. Versatile and open software for comparing large genomes. Genome Biol **5:**R12.

25. **Lackner, G., N. Moebius, L. Partida-Martinez, and C. Hertweck.** 2011. Complete Genome Sequence of Burkholderia rhizoxinica, an Endosymbiont of Rhizopus microsporus. J Bacteriol **193:**783-4.

26. **Lewenza, S., B. Conway, E. P. Greenberg, and P. A. Sokol.** 1999. Quorum sensing in Burkholderia cepacia: identification of the LuxRI homologs CepRI. J Bacteriol **181:**748-56.

27. **Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-9.

28. **Lipuma, J. J.** 2005. Update on the Burkholderia cepacia complex. Curr Opin Pulm Med **11:**528-33.

29. **Mahenthiralingam, E., A. Baldwin, and C. G. Dowson.** 2008. Burkholderia cepacia complex bacteria: opportunistic pathogens with important natural biology. J Appl Microbiol **104:**1539-51.

30. **Mahenthiralingam, E., T. A. Urban, and J. B. Goldberg.** 2005. The multifarious, multireplicon Burkholderia cepacia complex. Nat Rev Microbiol **3:**144-56.

31. **Marais, G. A., A. Calteau, and O. Tenaillon.** 2008. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica **134:**205-10.

32. **Nierman, W. C., D. DeShazer, H. S. Kim, H. Tettelin, K. E. Nelson, T. Feldblyum, R. L. Ulrich, C. M. Ronning, L. M. Brinkac, S. C. Daugherty, T. D. Davidsen, R. T. Deboy, G. Dimitrov, R. J. Dodson, A. S. Durkin, M. L. Gwinn, D. H. Haft, H. Khouri, J. F. Kolonay, R. Madupu, Y. Mohammoud, W. C. Nelson, D. Radune, C. M. Romero, S. Sarria, J. Selengut, C. Shamblin, S. A. Sullivan, O. White, Y. Yu, N. Zafar, L. Zhou, and C. M. Fraser.** 2004. Structural flexibility in the Burkholderia mallei genome. Proc Natl Acad Sci U S A **101:**14246-51.

33. **Peeters, E., A. Sass, E. Mahenthiralingam, H. Nelis, and T. Coenye.** 2010. Transcriptional response of Burkholderia cenocepacia J2315 sessile cells to treatments with high doses of hydrogen peroxide and sodium hypochlorite. BMC Genomics **11:**90.

34. **Romero-Gomez, M. P., M. I. Quiles-Melero, P. Pena Garcia, A. Gutierrez Altes, M. A. Garcia de Miguel, C. Jimenez, S. Valdezate, and J. A. Saez Nieto.** 2008. Outbreak of Burkholderia cepacia bacteremia caused by contaminated chlorhexidine in a hemodialysis unit. Infect Control Hosp Epidemiol **29:**377-8.

35. **Rose, H., A. Baldwin, C. G. Dowson, and E. Mahenthiralingam.** 2009. Biocide susceptibility of the Burkholderia cepacia complex. J Antimicrob Chemother **63:**502-10.

36. **Sajjan, U., C. Ackerley, and J. Forstner.** 2002. Interaction of cblA/adhesin-positive Burkholderia cepacia with squamous epithelium. Cell Microbiol **4:**73-86.

37. **Sajjan, U. S., L. Sun, R. Goldstein, and J. F. Forstner.** 1995. Cable (cbl) type II pili of cystic fibrosis-associated Burkholderia (Pseudomonas) cepacia: nucleotide sequence of the cblA major subunit pilin gene and novel morphology of the assembled appendage fibers. J Bacteriol **177:**1030-8.

38. **Sass, A., A. Marchbank, E. Tullis, J. J. LiPuma, and E. Mahenthiralingam.** Intrinsic, adaptive and evolutionary changes in the antibiotic resistance of Burkholderia cenocepacia observed by global gene expression analysis. BMC Genomics **Submitted**.

39. **Summer, E. J., C. F. Gonzalez, T. Carlisle, L. M. Mebane, A. M. Cass, C. G. Savva, J. LiPuma, and R. Young.** 2004. Burkholderia cenocepacia phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. J Mol Biol **340:**49-65.

40. **Tumapa, S., M. T. Holden, M. Vesaratchavest, V. Wuthiekanun, D. Limmathurotsakul, W. Chierakul, E. J. Feil, B. J. Currie, N. P. Day, W. C. Nierman, and S. J. Peacock.** 2008. Burkholderia pseudomallei genome plasticity associated with genomic island variation. BMC Genomics **9:**190.

41. **Vandamme, P., B. Holmes, T. Coenye, J. Goris, E. Mahenthiralingam, J. J. LiPuma, and J. R. Govan.** 2003. Burkholderia cenocepacia sp. nov.--a new twist to an old story. Res Microbiol **154:**91-6.

42. **Vanlaere, E., A. Baldwin, D. Gevers, D. Henry, E. De Brandt, J. J. LiPuma, E. Mahenthiralingam, D. P. Speert, C. Dowson, and P. Vandamme.** 2009. 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 **59:**102-11.

43. **Vanlaere, E., J. J. Lipuma, A. Baldwin, D. Henry, E. De Brandt, E. Mahenthiralingam, D. Speert, C. Dowson, and P. Vandamme.** 2008. 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 **58:**1580-90.

44. **Wernegreen, J. J.** 2005. For better or worse: genomic consequences of intracellular mutualism and parasitism. Curr Opin Genet Dev **15:**572-83.

45. **Yoder-Himes, D. R., P. S. Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, J. M. Tiedje, and R. Sorek.** 2009. Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. Proc Natl Acad Sci U S A **106:**3976-81.

46. **Zerbino, D. R., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821-9.

47. **Zlosnik, J. E., T. J. Hird, M. C. Fraenkel, L. M. Moreira, D. A. Henry, and D. P. Speert.** 2008. Differential mucoid exopolysaccharide production by members of the Burkholderia cepacia complex. J Clin Microbiol **46:**1470-3.

Chapter 5.
Thesis summary and outlook.

Rediscovering *Burkholderia* diversity via comparative genome analysis

The synergistic and often coupled advances in high-throughput sequencing, genomic applications, and bioinformatics during the past 10-15 years have helped propel microbial genomics into the 21st century and have provided many outstanding insights into microbial diversity, ecology and evolution. With a increasing number of even more powerful genomics technologies and tools in hand today, microbiologists are poised to continue to make significant advances in the fast paced field of microbial genomics. The work presented in this thesis has taken full advantage of recent-past and current technologies, along with the data derived from sequencing platforms and genome sequencing centers.

The *Burkholderia* represent an extremely interesting group of bacteria in terms of their phenotypic diversity, their ability to interact with many eukaryotic (and presumably) bacterial and archaeal organisms, their ability to survive and thrive in many environments, as well as their genetic diversity and genomic fluidity, which may provide them with a greater ability to adapt to fluctuations or perturbations in their local environment. While a number of *Burkholderia* species have been very well studied and characterized, in part due to their pathogenic interactions with humans, animals and plants, as well as their beneficial ones, a large number of new strains with novel functions continue to be discovered.

A number of *Burkholderia* strains and species have been the target of recent sequencing efforts. While the majority of *Burkholderia* genome projects center around biothreat and other pathogenic species, a growing number of strains with biotechnological applications or of agricultural importance have also recently been sequenced. This large number of projects has presented a great opportunity to explore this interesting group of

organisms in more detail. In addition, since several clades, with a number of species have been sequenced, this has presented an even greater opportunity to gain insight into the genomic evolution of these clades and the species therein. Furthermore, since a number of strains in several species have been completed, this allowed the testing and development of novel methods to study the genomes of organisms within a single genus, yet at different phylogenetic distances from one another.

Chapter 1 framed the comparative genomic analysis in the subsequent Chapters within the context of microbial diversity, and provided an overview of the *Burkholderia* and their many diverse interactions with their environment. Chapter 1 also supplied background on *Burkholderia* genomics projects, as well as discussed genomic sequencing technology and how next generation sequencing platforms are opening novel avenues of research thought impossible only a few years ago. The sheer volume of sequencing data was alluded to along with the need for novel computational and bioinformatic methods for dealing with this data deluge.

Chapter 2 tackles the question of genomic variability within the entire *Burkholderia* genus. While many phylogenetic trees of the many species of *Burkholderia* have often been, and continue to be reported, the 16S gene remains an important marker for such classification. However, it is clear that this conserved gene does not sufficiently discriminate between species and can often lead to conflicting branching patterns, depending on the type of analysis, the genomes used, the method of alignment, etc. We present in Chapter 2 a phylogeny based on 31 conserved genes, whose topology at major branches were further supported by a concatenated core gene tree. Broad scale similarity and pangenome analyses revealed the variable nature of the *Burkholderia* genome. Using

the methods outlined in the Chapter, over 55 thousand gene families were found among the 25 completed genomes. Interestingly, the core genus genome contains fewer than 850 gene families. Association of the core, variable and unique gene families with chromosome and function corroborated and further strengthened the results of several previous studies and the view that the main chromosome is under strong selective constraints, while the secondary replicons may allow *Burkholderia* the flexibility in sampling genetic material, either via lateral gene transfer or mutation.

An extension of this concept was taken for Chapter 3, meant in part as a perspective review together with novel data from four *Burkholderia* species. In attempting to define in terms of genetic content and variability several *Burkholderia* species, the question of how to define species, and thus how to select those genomes appropriate for a given analysis came to the fore. In addition, such selection criteria inevitably result in biases and interpretation of any analyses must be made carefully if the goal is to provide insights that may be broadly applicable to microbial genome evolution. While the subjects of study will certainly influence the interpretation of results, the methods used to obtain these results must also be approached with care (as with every experimental design). In Chapter 3, the application of both nucleotide and protein based methods were used to not only uncover the core and pangenomes of four *Burkholderia* species (each with 2-4 sequenced strains available), but also to highlight the utility of using whole genome alignments as a method to retrieve detailed comparative genomic data that can complement other approaches.

For Chapter 4, next generation sequencing data from the Illumina platform was analyzed. Given the speed with which next generation sequencing platforms can now produce sequencing data, we selected to generate data for a number of strains in a single

sequencing run over the course of only a week. Novel methods, borrowed and developed, were used to manage and store the data effectively, trim and perform quality control on the reads, map the sequences to a reference genome for ease of interpretation, and assemble the data *de novo*. Four clinical isolates of the ET-12 lineage were analyzed and compared with the recently published J2315 strain (2). While the analysis of these data is not entirely complete, given the unfinished status of these newly sequenced genomes, the methods used ensured robust reporting of differences when found. Interestingly, few novel sequences were found, yet a large number of mutations and a few deletions were uncovered, suggesting that perhaps this lineage, or the set of strains isolated in this one clinic, may be following an evolutionary trajectory toward genome degradation.

Despite the large number of completed and ongoing *Burkholderia* genome sequencing projects, we still need to fill in the gaps within the *Burkholderia* genome landscape (3). As we continue to explore the genomic diversity of the *Burkholderia* genus, as well as other microbes, new methods such as the ones described in this thesis will be required in order to analyze and make sense of the constantly increasing flood of sequence data. Finally, microbes do not live in isolation, but within a complex network of coexisting organisms (1), including archaea and eukaryotes. It is possible that by learning more about the interactions with *Burkholderia* partners, as well as understanding the partners themselves, we will come to comprehend more about the *Burkholderia*. In addition, the field of metagenomics opens the possibility to study *Burkholderia* within the context of a community and allowing us to peer into its population structure as well. Although novel methods will once again be required to decipher these data, this field holds great promise

for studying genome structure and adaptation. We are only beginning to enter these exciting times.

References

1. **Chaffron, S., H. Rehrauer, J. Pernthaler, and C. von Mering.** 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res **20:**947-59.

2. **Holden, M. T., H. M. Seth-Smith, L. C. Crossman, M. Sebaihia, S. D. Bentley, A. M. Cerdeno-Tarraga, N. R. Thomson, N. Bason, M. A. Quail, S. Sharp, I. Cherevach, C. Churcher, I. Goodhead, H. Hauser, N. Holroyd, K. Mungall, P. Scott, D. Walker, B. White, H. Rose, P. Iversen, D. Mil-Homens, E. P. Rocha, A. M. Fialho, A. Baldwin, C. Dowson, B. G. Barrell, J. R. Govan, P. Vandamme, C. A. Hart, E. Mahenthiralingam, and J. Parkhill.** 2009. The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol **191:**261-77.

3. **Williams, D., J. P. Gogarten, and P. Lapierre.** 2010. Filling the gaps in the genomic landscape. Genome Biol **11:**103.