

The Genus *Burkholderia*: Analysis of 56 Genomic Sequences

D.W. Ussery^a · K. Kiil^a · K. Lagesen^b · T. Sicheritz-Pontén^a ·
J. Bohlin^c · T.M. Wassenaar^{a,d}

^aCenter for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark; ^bDepartment of Informatics, University of Oslo, Blindern, Oslo, and the Centre for Molecular Biology and Neuroscience and Institute of Medical Microbiology, University of Oslo, Oslo, ^cNorwegian School of Veterinary Science, Oslo, Norway; ^dMolecular Microbiology and Genomics Consultants, Zotzenheim, Germany

Abstract

The genus *Burkholderia* consists of a number of very diverse species, both in terms of lifestyle (which varies from category B pathogens to apathogenic soil bacteria and plant colonizers) and their genetic contents. We have used 56 publicly available genomes to explore the genomic diversity within this genus, including genome sequences that are not completely finished, but are available from the NCBI database. Defining the pan- and core genomes of species results in insights in the conserved and variable fraction of genomes, and can verify (or question) historic, taxonomic groupings. We find only several hundred genes that are conserved across all *Burkholderia* genomes, whilst there are more than 40,000 gene families in the *Burkholderia* pan-genome. A BLAST matrix visualizes the fraction of conserved genes in pairwise comparisons. A BLAST atlas shows which genes are actually conserved in a number of genomes, located and visualized with reference to a chosen genome. Genomic islands are common in many *Burkholderia* genomes, and most of these can be readily visualized by DNA structural properties of the chromosome. Trees that are based on relatedness of gene family content yield different results depending on what genes are analyzed. Some of the differences can be explained by errors in incomplete genome sequences, but, as our data illustrate, the outcome of phylogenetic trees depends on the type of genes that are analyzed.

Copyright © 2009 S. Karger AG, Base

The genus *Burkholderia* belongs to the beta sub-division of Proteobacteria and contains a wide variety of Gram-negative species that occupy very different niches. Some are zoonotic pathogens, others are opportunistic human pathogens whilst yet others live harmless in the environment. Some species are able to degrade industrial waste compounds. Plant pathogens are also represented, and in contrast others protect plants against pathogens or promote plant growth. *Burkholderia* genomes consist of two or three chromosomes and frequently contain plasmids as well. Their genomes are large, variable, and extremely interesting as they can provide important insights to the evolutionary processes that shape bacterial genomes. The two species that attract attention

because of their potential in bio-terrorism are *B. mallei* and *B. pseudomallei*. With multiple genome sequences available for these species and for a number of related species, comparative genomics of the genus *Burkholderia* is now en vogue. Here we will compare 56 sequenced *Burkholderia* genomes and present observations to illustrate that presumed evolutionary relatedness depends on which fraction of the genome is analyzed. First, *B. mallei*, *B. pseudomallei* and the diseases they cause are introduced.

***Burkholderia mallei* Causes Glanders and *B. pseudomallei* Causes Melioidosis**

B. mallei is a nonmotile, nonsporulating, obligate aerobe organism previously known as *Pseudomonas mallei*. It causes glanders in horses and several other animal species. Animals contract the disease by ingestion of contaminated food or water. Traditionally, the disease is divided into nasal, pulmonary or cutaneous cases. The disease frequently progresses to septicaemia that will be fatal within days. A chronic form can occur in horses where nasal and subcutaneous nodules develop; such animals can be carriers for months or years before death occurs. The disease was once widespread, but by the mid-1900s it was eradicated in many countries by isolating and eradicating infected animals. It is still endemic in regions in Africa, Asia, the Middle East and Central and South America. A vaccine does not exist.

Human infections caused by *B. mallei* are rare although exceptionally few organisms are needed for human infection. Transmission from animal to man is inefficient and human-to-human spread is extremely rare. Cases result from direct and prolonged contact with infected domestic animals or from direct contamination with the infectious agent in the laboratory, presumably resulting from aerosols forming during routine handling. The low infectious dose, and the usual fatal outcome in humans, makes *B. mallei* a potential agent for biological warfare and bio-terrorism. Symptoms in humans depend on whether it is a localized cutaneous, pulmonary or bloodstream infection. Bloodstream infections have a fatality rate of 95% within a few days.

B. pseudomallei causes melioidosis, also known as Whitmore disease. The disease is similar to glanders but is restricted to the tropics and is endemic in tropical parts of Southeast Asia (notably Thailand), Australia and China. It is also found in tropical Africa and India. Occasionally, travelers import the disease into Europe or the US. In contrast to *B. mallei*, which is not frequently detected outside a host, *B. pseudomallei* survives in soil and water and it has a broader host range. As a consequence, human melioidosis is far more common than glanders and in some regions it accounts for 20 to 40% of community-acquired septicaemia. Melioidosis can be transmitted through contaminated water, notably during the rainy season, or by inhalation of contaminated dust. Human infections have a high mortality. The latent phase between infection and disease can be extremely long, up to months or even years and relapse is quite common.

B. mallei has most probably evolved from *B. pseudomallei*. This was concluded from multilocus sequence typing (MLST), a technique that assesses allelic variation in a

number of household genes [1]. In recognition of this close relationship, *B. pseudomallei* and *B. mallei* are both taxonomically included in what is called the Pseudomallei group.

Other *Burkholderia* Species Have a Variety of Lifestyles

In addition to *B. mallei* and *B. pseudomallei*, the genus *Burkholderia* contains more than 40 other species. Only those for which a genome sequence is available are listed here. Two of these belong to the Pseudomallei group: *B. thailandensis* also lives in tropical environments but is not pathogenic to mammals. *B. oklahomensis* has been described as ‘*B. pseudomallei*-like’, but MLST and DNA-DNA hybridization have identified it as a novel species [2]. *B. oklahomensis* has been isolated from wounds associated with soil contamination.

Another important group of closely related species is the *B. cepacia* complex (BCC), wherein each species is also known as a genomovar, with *B. cepacia* as genomovar I. (There are more than nine species within BCC, with recent novel additions [3], but their genomes have not yet been sequenced). They are all opportunistic pathogens, frequently causing infections in cystic fibrosis patients where the infection can be fatal. Besides this relevance to human medicine, a number of species of the BCC also have other interesting properties. *B. cenocepacia* (genomovar III) is ubiquitous in the environment as a phytopathogen. *B. dolosa* was formerly known as *B. cepacia* genomovar IV. *B. multivorans* cannot transmit from patient to patient, in contrast to the other BCC species. *B. ambifaria* (genomovar VII) has attracted interest since it lives in the rhizosphere of pea plants where it can protect the plants against pathogens. *B. vietnamiensis* is also beneficial to plants and has been studied as a growth-promoting bacterium. It has also bioremediation properties as it can degrade aromatic hydrocarbons such as benzene and toluene. *B. ubonensis* (also known as *B. uboniae*) is a common soil bacterium that is proposed as a new member of the BCC [4]. The latest addition of the BCC for which a genome sequence is available is *B. lata*, first described in 2009 [5].

The remainder of species for which a genome species is available are not pathogenic to humans and do not belong to a particular subgroup. *B. xenovorans* is an environmental organism of economic importance as it can degrade polychlorinated biphenyl (PCB) compounds. In contrast, *B. phymatum* lives in symbiotic relationship with tropical legumes. *B. phytofirmans* is also beneficial to its plant host, and lives outside the tropics. *B. graminis* is found in the rhizosphere of Gramineae plants, such as wheat and corn.

The First *Burkholderia* Genome Sequences

The potential use in biological warfare raised a scientific interest that resulted in a relatively large number of published genome sequences. The genome of *B. mallei*

contains two chromosomes and the first complete sequence was published in 2004 (*B. mallei* strain ATCC 23344) [6]. At the same time the sequence for both chromosomes of *B. pseudomallei* strain K96243 was published [7]. A large number of insertion sequences were found in the *B. mallei* genome that have mediated multiple deletions and rearrangements compared to the genome of *B. pseudomallei*. The genome of the latter contained 16 genomic islands that appeared absent in the smaller genome of *B. mallei*. The authors speculated that these genomic islands had been absent from the genetic repertoire of the *B. pseudomallei* ancestral clone that produced *B. mallei* [7]. Gene loss would be consistent with the reduced adaptive potential and restricted host specificity of *B. mallei* compared to *B. pseudomallei*. Other differences between the two species observed related to the fact that *B. pseudomallei* is motile but *B. mallei* is not (a few of its motility genes have undergone mutations as a result of release of selective pressure), and that *B. pseudomallei* can secrete a number of toxins that *B. mallei* produces but cannot secrete, due to a mismatch in a secretory system component. Finally, the *B. mallei* genome contains two type III secretion systems on chromosome 2, which contributes to its virulence potential.

The two species share an exceptionally high number of local direct repeat sequences, covering more than 20% of the total length of the chromosomes. We classify repeats as 'local' when they are found by searching with a 15 nucleotide (nt) window within a 100 nt region, and as 'global' when determining the frequency of 100 nt-long sequences repeated anywhere on the genome [8]. The two chromosomes of each species also showed significant functional partitioning, with the large chromosome 1 (4.1 Mb in *B. pseudomallei*, 3.5 Mb in *B. mallei*) encoding many genes involved in metabolism and growth, the smaller chromosome 2 (3.2 Mb and 2.3 Mb, respectively) containing genes related to adaptation and survival in different niches.

The genome of *B. thailandensis* was sequenced in 2006 but already in 2004 it was recognized that its genome had also undergone gene reduction compared to *B. pseudomallei* [9]. This work was based on microarray analysis using partial genome sequences of *B. pseudomallei* K96243. The authors concluded that genome reduction of *B. thailandensis* occurred independent of that of *B. mallei*, possibly by different mechanisms, as the deleted genes were not found present in clusters in *B. pseudomallei*, but rather dispersed over its genome. When the *B. thailandensis* genome sequence became available, it was obviously compared to *B. pseudomallei* [10]. The authors concentrated on *B. mallei* genes that are up- or downregulated during colonization in a mouse model, and found that down-regulated genes were more strongly conserved in *B. thailandensis* than in *B. pseudomallei*.

Over time more *Burkholderia* genome sequences have been finished, such as that of *B. xenovorans* LB400 [11]. Its genome contains three chromosomes, totaling 9.73 Mb, though other strains can have smaller genomes with 7.4 Mb being the currently known minimum. As in the other *Burkholderia* species, the chromosomes have undergone functional specialization and the two smaller chromosomes have undergone less selective pressure, allowing for more variation. As the number of genome sequences

grew, including multiple genomes for a number of species, the comparison within and between species became truly interesting. A database especially dedicated to *Burkholderia* genomes has recently been established at www.burkholderia.com [12].

Genome sequences do not have to be complete (with each chromosome in a single, contiguous piece) to be used for comparative analysis. Incomplete genome sequences are frequently released into the public domain as multiple contigs, and sometimes it is left to that. Here we perform comparative genomic analysis of partial and complete genome sequences within the *Burkholderia* genus that are publicly available.

Practicalities of Large-Scale Comparative Genomics: Introducing the BLAST Matrix

The 56 *Burkholderia* genome sequences available at the time of writing are summarized in table 1. The number of contigs is given for all genomes. Working with such large number of genomes one can soon be overwhelmed with data: the interpretation and graphical representation of findings becomes a real issue. We largely concentrate on coding regions, and here we zoom in on the degree of gene conservation between genomes, ignoring gene location, chromosome separation or gene synteny. We did not perform a detailed analysis of gene function, nor did we relate individual genes to the characteristics of that particular strain or species (thus respecting the objectives of any sequencing project). This simplified approach allowed us to do large-scale analysis of gene conservation and chromosome evolutionary processes.

The approach is quite straightforward: Starting with one chromosome as a query, every gene is compared by BLAST to a second genome and conserved genes are scored. After all genes of the query genome are checked, the next genome is chosen to compare with the query genome until all genomes have been screened. Then the next genome is used as a query source, again checking all its individual genes against all other genomes. This way every genome in the analysis set will serve as a query against all others, and will also be queried by all other genomes [8].

Comparison of amino acid sequences of coding regions requires a standardized gene finding process, in order to rule out differences introduced by various (automated) gene identification programs. Genomes are frequently over- or under-annotated and occasionally the wrong strand of a gene is annotated [13]. Over-annotation is frequently seen in very short open reading frames, which can be erroneously recognized as genes if the cut-off for gene finding is taken too low (although some very short open reading frames can indeed be true genes). Under-annotation is sometimes observed for non-translated genes, such as tRNA or even rRNA genes that can be missing in a genome annotation. In our analysis only amino acid sequences were used, and non-translated RNA genes were excluded. In order to avoid artificial variation in our analysis, all used *Burkholderia* genomes were annotated by a standard gene finding and annotation program, so that arbitrarily chosen cut-offs would be consistent and not influence comparative analyses [14, 15].

Table 1. Genome sequences included in this study. All genomes used are publicly available for analysis

Group	Species	Strain ^a	No. of contigs ^b	PID	Sequence Source ^c
Pseudomallei group	<i>B. pseudomallei</i>	1106a	2	16182	TIGR
	<i>B. pseudomallei</i>	1710b	2	13954	TIGR
	<i>B. pseudomallei</i>	668	2	13953	TIGR
	<i>B. pseudomallei</i>	K96243	2	178	Sanger Institute
	<i>B. pseudomallei</i>	576	21	31091	LANL
	<i>B. pseudomallei</i>	305	36	18775	TIGR
	<i>B. pseudomallei</i>	S13	169	13951	TIGR
	<i>B. pseudomallei</i>	1655	194	13949	TIGR
	<i>B. pseudomallei</i>	1106b	202	16181	TIGR
	<i>B. pseudomallei</i>	1710a	209	13950	TIGR
	<i>B. pseudomallei</i>	Pasteur 52237	217	13952	TIGR
	<i>B. pseudomallei</i>	406e	271	16231	TIGR
	<i>B. pseudomallei</i>	BCC215	1030	19491	NMRC
	<i>B. pseudomallei</i>	NCTC 13177 (WKO97)	1077	19493	NMRC
	<i>B. pseudomallei</i>	112	1274	19495	NMRC
	<i>B. pseudomallei</i>	B7210	1424	19499	NMRC
	<i>B. pseudomallei</i>	7894	1568	19497	NMRC
	<i>B. pseudomallei</i>	91	1690	19505	NMRC
	<i>B. pseudomallei</i>	9	1762	19503	NMRC
	<i>B. pseudomallei</i>	14	1888	19507	NMRC
	<i>B. pseudomallei</i>	DM98 (BCC11)	2371	19509	NMRC
	<i>B. mallei</i>	ATCC 23344	2	171	TIGR
	<i>B. mallei</i>	NCTC 10229	2	13943	TIGR
	<i>B. mallei</i>	NCTC 10247	2	13946	TIGR
	<i>B. mallei</i>	SAVP1	2	13947	TIGR
	<i>B. mallei</i>	ATCC 10399	106	13944	TIGR
	<i>B. mallei</i>	GB8 horse 4	181	13945	TIGR
	<i>B. mallei</i>	JHU	184	13988	TIGR
	<i>B. mallei</i>	FMH	205	13987	TIGR
	<i>B. mallei</i>	2002721280	208	16352	TIGR
	<i>B. mallei</i>	PRL-20	272	19147	TIGR
	<i>B. thailandensis</i>	E264 ^d (ATCC 700388)	2	10774	TIGR
	<i>B. thailandensis</i>	Bt4	803	19533	NMRC
	<i>B. thailandensis</i>	TXDOH	810	19541	NMRC
<i>B. thailandensis</i>	MSMB43	1230	19501	NMRC	
<i>B. oklahomensis</i>	C6786 ^d	633	19535	NMRC	
<i>B. oklahomensis</i>	EO147	886	19537	NMRC	
Complex (BCC)	<i>B. cenocepacia</i>	J2315	4	339	Sanger Institute
	<i>B. cenocepacia</i>	AU 1054	3	13919	DOE
	<i>B. cenocepacia</i>	H12424	4	13918	DOE
	<i>B. cenocepacia</i>	MC0-3	3	17929	DOE
	<i>B. cenocepacia</i>	PC184	174	16169	Broad Institute

Table 1. Continued

Group	Species	Strain ^a	No. of contigs ^b	PID	Sequence Source ^c
	<i>B. multivorans</i>	ATCC 17616	4	17407	DOE
	<i>B. ambifaria</i>	AMMD ^d	4	13490	DOE
	<i>B. ambifaria</i>	MC40-6	4	17411	DOE
	<i>B. ambifaria</i>	IOP40-10	629	20669	DOE
	<i>B. ambifaria</i>	MEX-5	706	20667	DOE
	<i>B. dolosa</i>	AU0158	233	16168	Broad Institute
	<i>B. vietnamiensis</i>	G4	8	10696	DOE
	<i>B. ubonensis</i>	Bu	1143	19539	NMRC
	<i>B. lata</i>	383	3	10695	DOE
None	<i>B. phymatum</i>	STM815	4	17409	DOE
None	<i>B. phytofirmans</i>	PsJN ^d	3	17463	DOE
None	<i>B. xenovorans</i>	LB400	3	254	DOE
None	<i>B. graminis</i>	C4D1M ^d	70	20537	DOE
None	<i>Burkholderia spp.</i>	H160	310	29197	DOE

^a Alternative names appear between parentheses.

^b Number of contigs below 10 indicate that all chromosomes and plasmids are in one piece.

^c DOE = US Department of Energy Joint Genome Institute; TIGR = The Institute of Genome Research; NMRC = Naval Medical Research Center/Defense Research Directorate, Genomics, USA. LANL = Los Alamos National Laboratory. Inst = Institute.

^d Type strain of the species.

Another difficulty of comparisons of coding sequences is to decide when to call a pair of genes ‘conserved’. This balancing act has two opposing risks. One can set very strict rules of identity, so that genes have to be highly similar in order to be screened as ‘conserved’ (in gene sequence and thus presumably in biological function). Consequentially, this may result in a very high number of genes without homologs, which decreases the significance of the findings. Alternatively, one can set relatively loose requirements for conservation, but then genes may be grouped together that have different biological functions as a result of divergent evolutionary processes, which also results in questionable results. As a rule-of-thumb, we have found that two genes need to have at least 50% identity over at least 50% of their lengths in order to be scored as conserved. This 50–50 rule has been found satisfactory for a number of species and genera that we analyzed. By varying these parameters (for instance 40% identity over at least 70% of sequence length) we observed that the analysis was quite robust.

The next challenge faced is how to represent the findings. BLAST produces long lists summarizing the findings that are obviously not conceivable or interpretable in their raw form. The data were instead condensed to two numbers per genome, indicating how many genes were tested as query and what fraction of these found

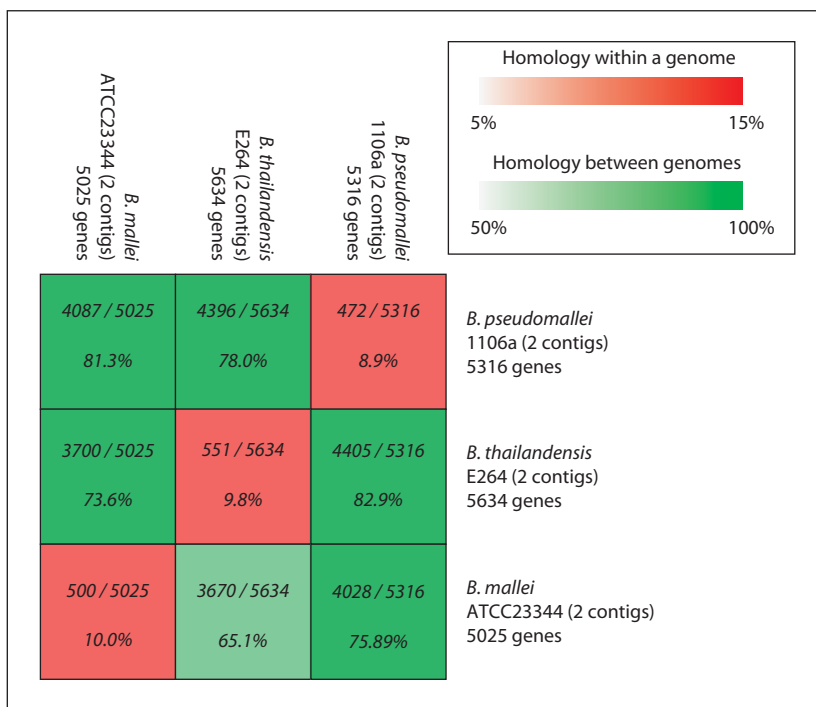


Fig. 1. BLAST matrix of *Burkholderia* genomes of three species. The scores in each field give the number of homologous genes per number of total genes in the tested genome, followed by percentage. The coloring of the cells depends on this fraction. The red cells represent homologous genes detected within one genome. The color scales can be adjusted according to the spread of the percentages in the analyzed genomes.

homologs in the blasted genome. These numbers can be shown in a matrix [16] of which figure 1 shows a simplified example.

The cells of the matrix are colored according to the fractions of homology: the higher this percentage, the more intense a color is used. In this way even very large BLAST comparisons can still be captured in a figure that immediately reveals its information by visual inspection. An example is given in figure 2, where 28 genomes are compared of 4 *B. mallei*, 4 *B. thailandensis* and 20 *B. pseudomallei* strains. For this matrix the color scale has been adjusted to cover a wider range. From this matrix it is obvious (even without being able to read the actual numbers) that 9 *B. pseudomallei* genomes form a group within this species, and these are less homologous to the others, indicated by the lighter color of the matrix cells. The four *B. mallei* genomes are quite similar, as they report similar homology percentages (similar color intensities) for all comparisons. In contrast, the four *B. thailandensis* genomes differ considerably. It should be noted, however, that the *B. thailandensis* genome indicated by the arrow still consists of >1200 contigs; this indicates its sequence is still incomplete, and that may explain why fewer homologous genes are detected in this genome.



Fig. 2. BLAST matrix of 28 *Burkholderia* genomes, belonging to 4 *B. mallei*, 4 *B. thailandensis* and 20 *B. pseudomallei* strains. The arrow identifies the *B. thailandensis* MSMB43 genome whose sequence is still relatively incomplete.

Zooming in at Genes: Comparing Genomes in a BLAST Atlas

Although a BLAST matrix as shown in figure 2 gives valuable insights into which genomes are more and which are less closely related, it only reports information on the number of homologous genes. The matrix does not contain information about the identity of these genes, or whether the same set of genes is conserved in the next pairwise alignment. To capture such data, an atlas is more suitable [17].

Figure 3 shows a Genome Atlas of *B. cenocepacia* strain J2315, for all three chromosomes and the plasmid. Although the sequence had been finished a few years ago, it has only recently been published [18]. Three lanes have been added to a classical Genome

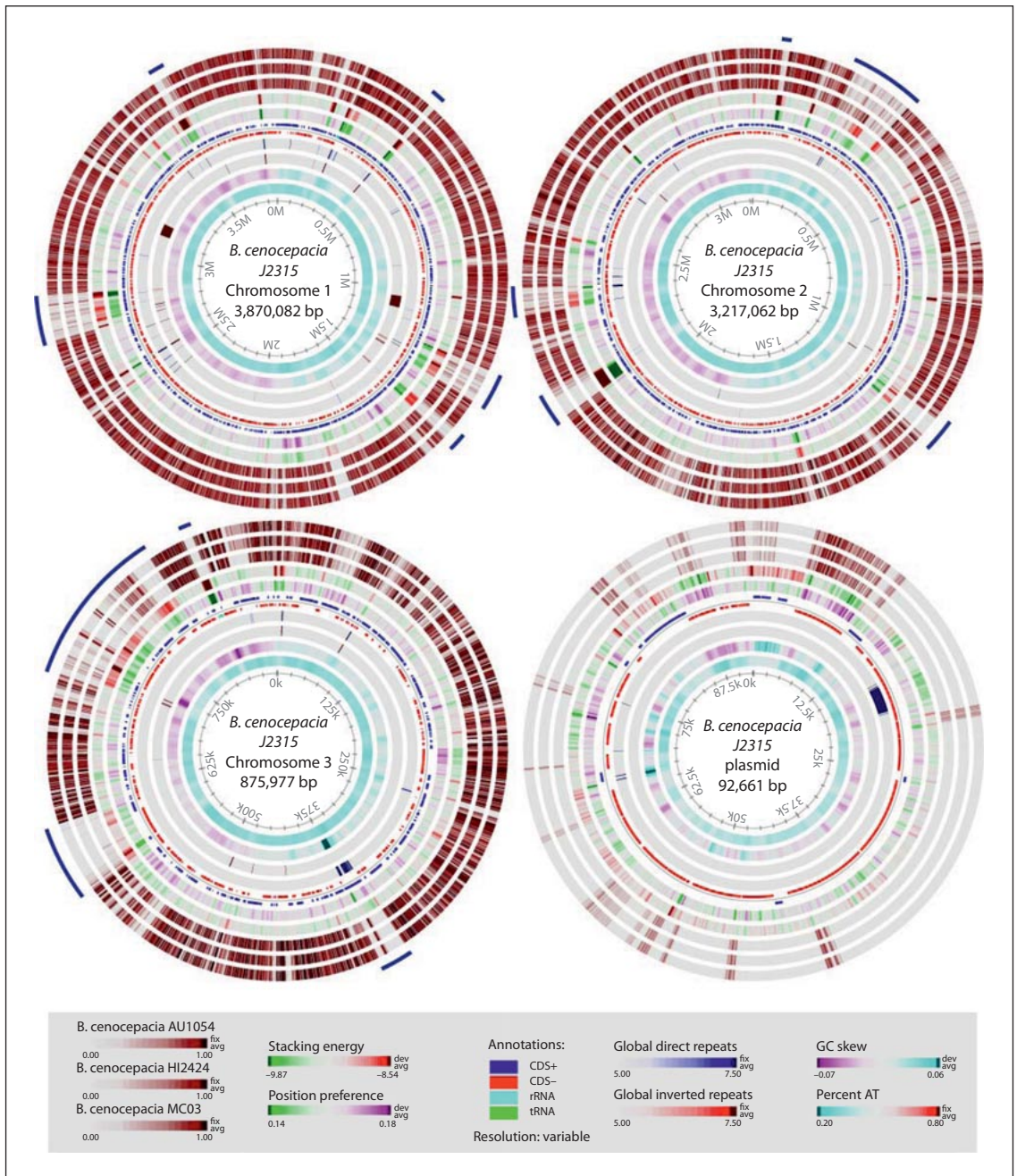


Fig. 3. Genome Atlases for the genome of *B. cenocepacia* strain J2315, with three BLAST lanes added for other *B. cenocepacia* genomes. The scale of the three chromosomes and the plasmid obviously differ. The location of genome islands present in J2315, recognizable by DNA structural properties and by their absence in the other genomes, is indicated by blocks around each chromosomal atlas.

Atlas (as already introduced in the first chapter of this book [19]): the outer three lanes show which genes of the J2315 genome are conserved (as identified by BLAST) in other sequenced *B. cenocepacia* strains. The figure illustrates that the largest chromosome is the most conserved of the four DNA entities, and that the plasmid is the least conserved. The BLAST lanes identify regions in the J2315 chromosomes that are not conserved in the other *B. cenocepacia* genomes. Some of these regions (marked in fig. 3) also report DNA structural properties that are unique from the rest of the chromosomes, and these happen to be the genomic islands for strain J2315. Genes present in the plasmid of strain J2315 are not found in the other three strains, except for a locus around 4–10 kb, which contains a few genes including a DNA polymerase III subunit. This kind of analysis does not reveal whether the BLAST matches are also plasmid-encoded in the other strains; in fact, neither *B. cenocepacia* AU1054 nor MC03 do carry plasmids.

Given that genomic islands are frequent in *Burkholderia* genomes [20], and most of these are species or even isolate-specific, we asked the question whether the species or even the genus can still be considered as a more-or-less uniform group, to which the concept of an evolutionary tree would still hold.

The Pan- and Core Genomes of *Burkholderia* Species

Figure 3 identifies which genes that are present in one particular *Burkholderia* genome are conserved in other genomes of the species. Such analysis can be extended to identify the fraction of genes that is always present in every *Burkholderia* genome, which we call the core genome of the genus. (A core genome was previously introduced with a less strict definition to comprise genes that are present in most individuals [21], but we use here a stricter definition). The conserved core genome can be determined for a genus or a species, provided sufficient genome sequences are available, and the sequenced strains truly represent the diversity that is out there. A core genome will decrease in size as more genomes are added, as genes that were found conserved in one lot of genomes may be lacking in a next added genome. Eventually, the curve will flatten out if the true number of conserved genes is reached.

Together with the core genome, a pan-genome can be defined, which represents all genes potentially present in a genome of a particular species or genus. The concept of a pan-genome was first introduced by Tettelin and coworkers who compared 8 different *Streptococcus agalactiae* genomes [22]. Genes or gene families that are not part of the core genome are called ‘accessory’ or ‘auxiliary’. The pan-genome will increase with each added genome, as novel genes are discovered for each added genome. Again, this curve is expected to flatten out when the true pan-genome of a species (genus) is covered. More about pan- and core genomes is described in [8].

When the pan- and core genomes of one species (say, *B. pseudomallei*) have thus been established, a genome of a different species could be added, say a *B. mallei*, to see what effect this new species has to the pan- and core genome curves. This is illustrated

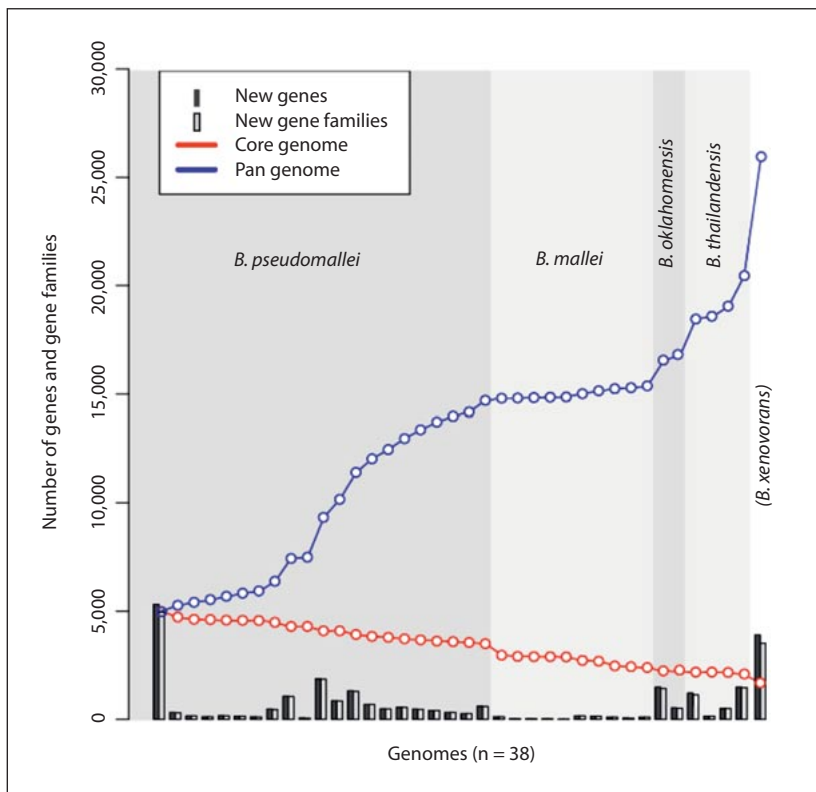


Fig. 4. Pan- and core genome plot of the Pseudomallei group currently consisting of 21 *B. pseudomallei*, 10 *B. mallei*, 4 *B. thailandensis* and 2 *B. oklahomensis* genomes. A *B. xenovorans* genome is added at the end for comparison. Within the species, the genomes are ordered for increasing numbers of genes.

in figure 4, where the Pseudomallei group is analyzed. As can be seen, the pan-genome curve for *B. pseudomallei* does not yet reach a plateau after 21 genomes; apparently, the true diversity of this species has not yet been covered. Compared to this, the curves of *B. mallei* are much more flattened, indicating less genetic diversity within this species. Note the drop in the core genome curve when leaving *B. pseudomallei* and entering *B. mallei*. This drop is caused by genes conserved in *B. pseudomallei* but not in *B. mallei*. Addition of the two *B. oklahomensis* genomes and after that the four *B. thailandensis* genomes adds quite a few genes to the pan-genome but hardly influences the core genome. In contrast, addition of *B. xenovorans* (which does not belong to the Pseudomallei group) causes a significant increase in the pan-genome and drop in the core-genome curve. This illustrates how far removed *B. xenovorans* is from the Pseudomallei group, in terms of the fraction of shared genes. Plots like these can thus assess the relatedness of isolates within and between taxonomic divisions.

From figure 4 we can see that the core genome of *B. pseudomallei* covers only approximately 4,000 of the 5,000 genes or gene families (80%) in a single genome

whereas the pan-genome easily comprises 15,000 genes (remember that the pan-genome is an artificial sum of all genes encountered in the analyzed genomes and by far exceeds the number of genes in a single genome). For *B. mallei*, the core genome comprises approximately 58% (2,800 genes out of 4,800) of a small *B. mallei* genome (this cannot be read from figure 4 as *B. mallei* is not the first species listed here). In an experimental approach based on micro-array analysis, the conserved gene fraction of *B. pseudomallei* was estimated in the same order as our estimated core genome, as 85% [23]. Their findings pointed out that human clinical isolates of *B. pseudomallei* clustered together on a tree based on the variable gene content. This suggests that virulence potential is largely coded in the variable gene fraction and as a consequence not all *B. pseudomallei* isolates would be equally virulent. The results presented here illustrate how a pan- and core genome analysis can identify genes of interest for pathogenicity research. The beauty of this analysis is that it identifies which genes belong to the variable fraction of a genome, so that a detailed analysis of their functions and interrelationships can easily follow. Pan- and core genome analysis is a promising strategy to include in the field of pathogenomics.

Figure 5 represents the pan- and core genome of the *Burkholderia* genus, extracted from all currently sequenced genomes. The figure shows that the pan-genome of the genus *Burkholderia* contains over 40,000 gene families, which is more than the number of genes present in a human genome. The large number of gene families of this genus is most likely due to the enormous diversity within this genus. The core genome of the genus, however, has decreased to only a few hundred genes that are conserved across all *Burkholderia* genomes.

Phylogenetic Trees

One simple analysis to perform for any complete or incomplete genome is to extract the 16S rRNA (*rrn*) gene(s) and to produce a tree including related isolates or species, as this can be used as confirmation that the correct DNA was sequenced. Examples of the 'wrong' organism being sequenced exist, and can arise from contamination during cultivation, DNA extraction, cloning and sequencing or even due to contamination (overwriting) of sequencing files. Incomplete genome sequences do not always include the *rrn* genes, as these are often repeated on a chromosome, and such repeats complicate the assembly process, so that they are temporarily removed from the raw sequences.

Figure 6 shows a phylogenetic tree based on 16S rRNA extracted from 56 genomes. As expected, there is little resolution within a species, due to the high degree of similarity of the 16S rRNA sequences from the same species. In light of the assumed ancestry of *B. mallei*, it is not surprising that the *B. pseudomallei* and *B. mallei* genes are somewhat mixed up, as nearly all of these are very similar (the long branch of *B. pseudomallei* 305 is probably an artefact due to a sequencing error, as this genome is not finished yet), and they are clearly separated from the BCC group (which are all

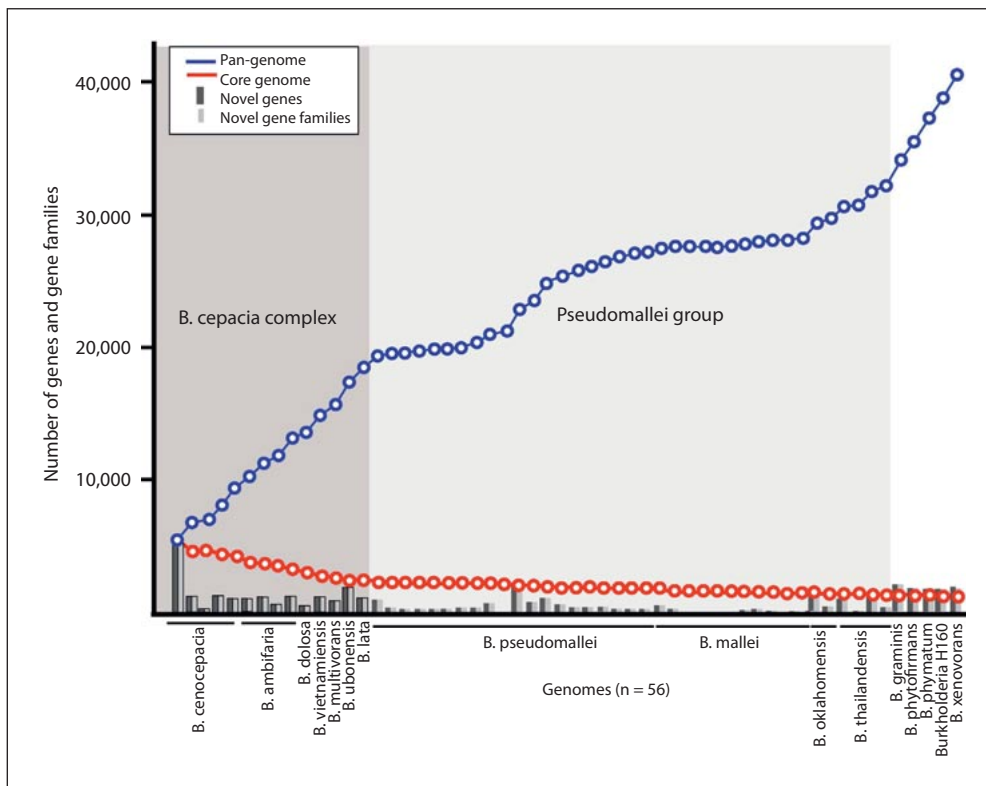


Fig. 5. Pan- and core genome plot of all 56 genome sequences from table 1, sorted for group and species. The BCC complex is plotted first, followed by the Pseudomallei group and last the species that do not belong to any group.

depicted in shades of blue). However, the *B. thailandensis* 16S rRNA genes are positioned as outliers of the Pseudomallei group, and one of them is somewhat in between that and the BCC group (indicated by an arrow). Moreover, the two *B. oklahomensis* 16S rRNA genes do not cluster within the Pseudomallei group, where they would be if their ‘Pseudomallei-like’ nature was reflected by their 16S rRNA. Finally, *B. ubonensis* is an outlier, and not positioned within the BCC group where it was reported previously [24]. Note, however, that the *rrn* sequence was extracted from a rather premature genome sequence (it was still in 1143 contigs) so it may still contain sequencing errors. Matching our expectations are *B. xenovorans*, *B. phytofirmans* and *B. phymatum* that are only distantly related to the other species. The unspecified genome, of isolate H160, has a ribosomal gene quite different to all other *Burkholderia* genes analyzed.

The method of MLST is used to analyze population genetics within a species, or between members of closely related species. For *Burkholderia*, partial sequences of 7 genes are usually analyzed but different schemes exist [25, 26]. We extracted the DNA fragments described in reference 24 from the genomes and analyzed these as one

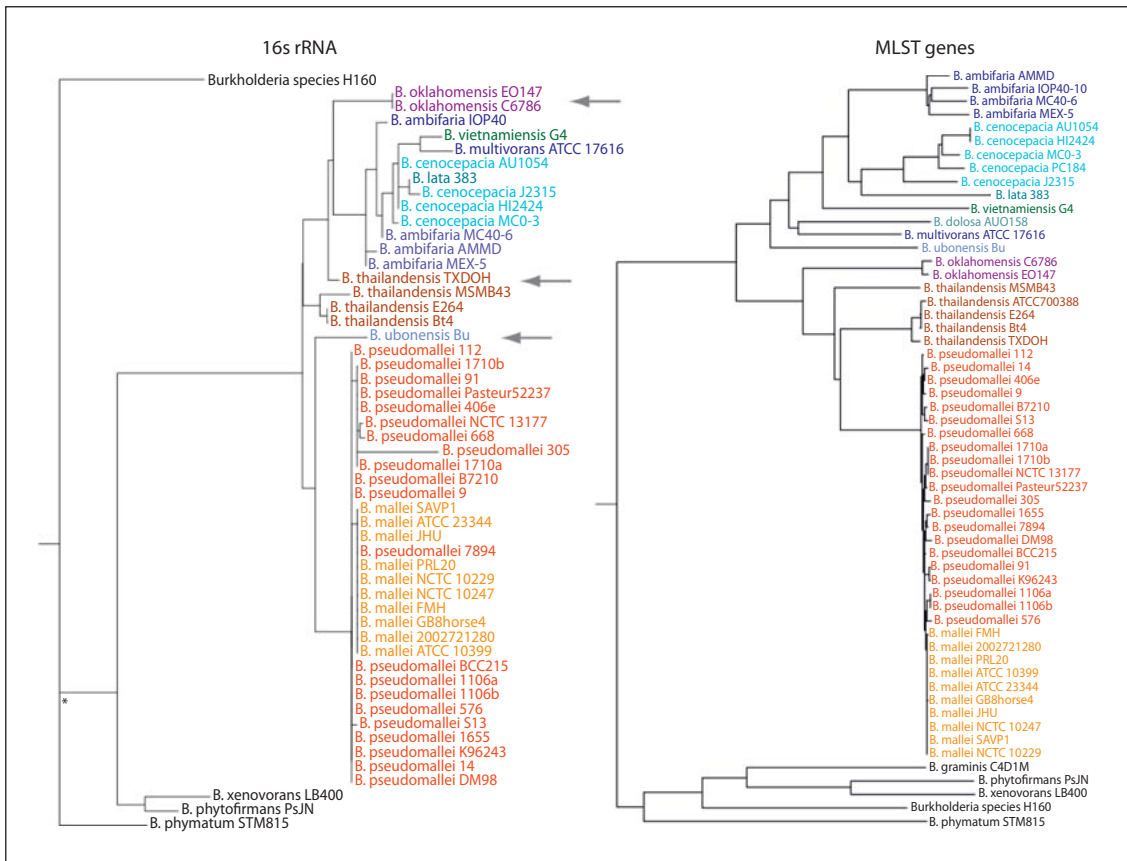


Fig. 6. To the left: a phylogenetic tree of the 16S rRNA gene (*rrn*) extracted from 53 genome sequences. One gene per genome was analyzed. *B. cenocepacia* PC184, *B. graminis* and *B. dolosa* were excluded, due to the lack of a full length 16S rRNA gene in these partially sequenced genomes. Genomes are color-coded according to species. Grey arrows indicate genes positioned different from expectations. The node for *B. phymatum* produced low bootstrap values (<500/1,000), indicated by an asterisk. To the right: phylogenetic tree of 7 concatenated MLST genes [24] extracted from 56 genomes.

artificially concatenated piece. This produced a tree (by neighbor joining) as shown to the right of figure 6. In this tree all proposed members of the Pseudomallei group cluster together with *B. thailandensis* and *B. oklahomensis* as closely related, and all members of the BCC group cluster as well. So this tree, based on all MLST genes combined, matches the currently used grouping better than the tree based on the *rrn* gene. *Burkholderia* species H160 could not be analyzed as its MLST genes were not yet completely sequenced.

Would the addition of more genes produce a similar tree? After all, MLST genes are supposed to be marker genes for the genetic relationship of most of the genome. The problem is that genes can be exchanged between (and within) species by horizontal gene transfer, so that they no longer produce consistent trees. To get around

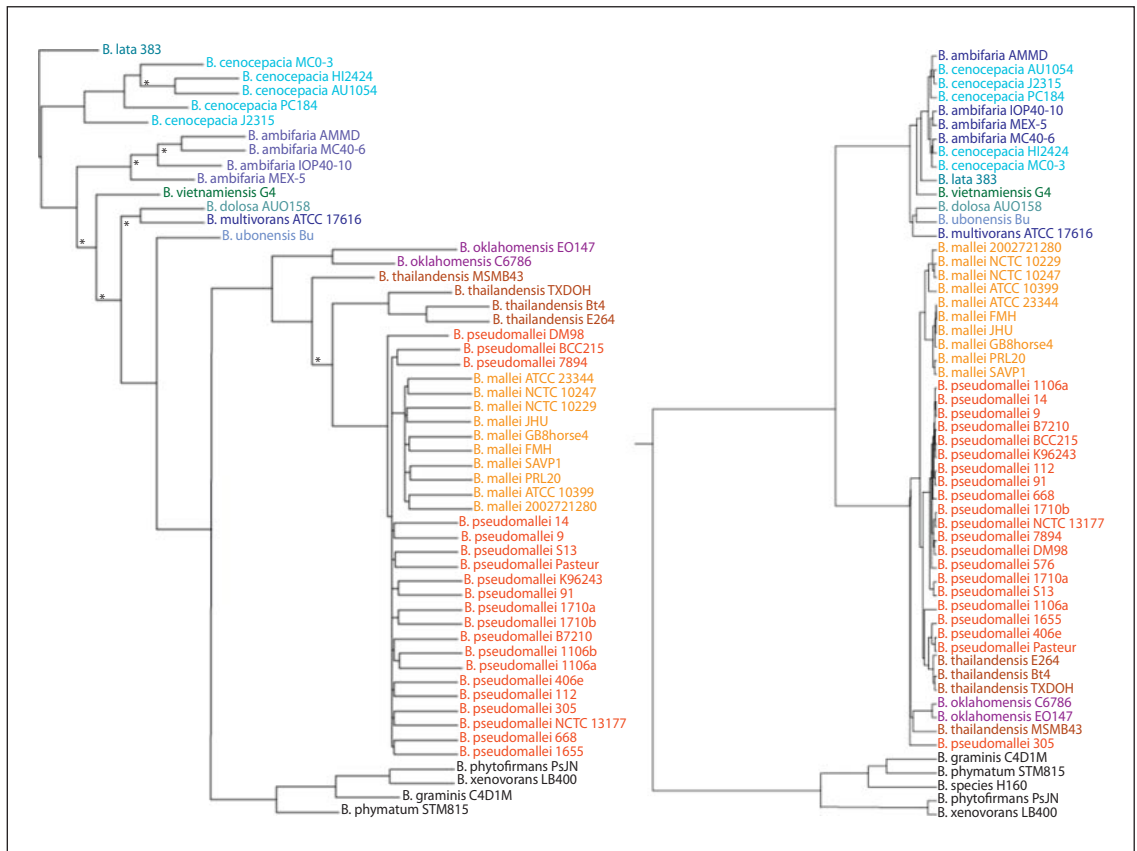


Fig. 7. The tree on the left is based on 612 protein genes that gave consistent trees when individually analyzed. Bootstrap values below 50/100 are indicated with an asterisk. The clustering on the right is based on the observed frequency of tetranucleotides compared to expected values, using a first-order Markov chain model. Such a clustering is independent of genes.

this, we identified those genes that produce consistent trees, so as to concentrate on genes to be least influenced by horizontal gene transfer. The tree to the left of figure 7 is based on 612 genes that are part of the *Burkholderia* core genome and produced consistent trees. Note that this is only about 12–15% of all genes in a given genome. The tree clearly separates the BCC group, the Pseudomallei group and those species not dedicated to any group. The biggest difference between the tree in figure 7 and the MLST tree in figure 6 is that the genomes now produce branches within a species, as there is more intra-species variation between 612 genes than between 7 (MLST) genes. We believe that figure 7 is a more complete representation of the true similarity and differences of these investigated organisms than the MLST tree provides.

All analyses presented so far concentrated on RNA or protein-coding genes, but it is also possible to compare the complete DNA sequence of the genome, irrespective

of what the nucleotides code for. One way to do so is to compare the frequency of oligomers, such as tetranucleotides, and compare this distribution to statistically expected values. The latter can be calculated in various ways, for example based on a first-order Markov chain model. The result is a genomic signature that is likely to be reflective of an organism's environment, as well as reflective of relatedness [27]. This 'genomic signature' is not affected by the number of contigs of a genome sequence and is independent of where on the genome it is searched for. The panel to the right of figure 7 shows such a clustering, and in general the observed arrangement is in agreement with the groupings of the other trees. It is reassuring that two completely independent methods result in similar clusters, and this suggests that these groupings are a true reflection of biological relationship.

In summary, we find that determining the taxonomic grouping of several of the *Burkholderia* species, based on their genomic sequences, is possible, but we suggest not to base this on a single (as in *rrn* analysis) or a few (as in MLST) genes, but rather to analyze a large number of genes or the complete DNA sequence, in order to optimally reflect the true genetic relationship between organisms. With the number of bacterial genome sequences steadily increasing, this approach will become more and more applicable to other species as well.

Acknowledgement

We thank the several sequencing centers that have deposited unfinished genomic data into the RefSeq database at NCBI. In particular, we would like to thank Tim Reed for kindly providing us with permission to use the as yet unpublished sequences of 15 *Burkholderia* genomes.

References

- 1 Godoy D, Randle G, Simpson AJ, Aanensen DM, Pitt TL, et al: Multilocus sequence typing and evolutionary relationships among the causative agent of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J Clin Microbiol* 2003; 41:2068–2079.
- 2 Glass MB, Steigerwalt AG, Jordan JG, Wilkins PP, Gee JE: *Burkholderia oklahomensis* sp. nov., a *Burkholderia pseudomallei*-like species formerly known as the Oklahoma strain of *Pseudomonas pseudomallei*. *Int J Syst Evol Microbiol* 2006;56:2171–2176.
- 3 Vanlaere E, Lipuma JJ, Baldwin A, Henry D, De Brandt E, et al: *Burkholderia latens* sp. nov., *Burkholderia diffusa* sp. nov., *Burkholderia arboris* sp. nov., *Burkholderia seminalis* sp. nov. and *Burkholderia metallica* sp. nov., novel species within the *Burkholderia cepacia* complex. *Int J Syst Evol Microbiol* 2008;58:1580–1590.
- 4 Yabuuchi E, Kawamura Y, Ezaki T, Ikedo M, Dejsirilert S, et al: *Burkholderia uboniae* sp. nov., L-arabinose-assimilating but different from *Burkholderia thailandensis* and *Burkholderia vietnamiensis*. *Microbiol Immunol* 2000;44:307–317.
- 5 Vanlaere E, Baldwin A, Gevers D, Henry D, De Brandt E, et al: Taxon K, a complex within the *Burkholderia cepacia* complex, comprises at least two novel species, *Burkholderia contaminans* sp. nov. and *Burkholderia lata* sp. nov. *Int J Syst Evol Microbiol* 2009;59:102–111.
- 6 Nierman WC, DeShazer D, Kim HS, Tettelin H, Nelson KE, et al: Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci USA* 2004;101:14246–14251.

- 7 Holden MT, Titball RW, Peacock SJ, Cerdeño-Tárraga AM, Atkins T, et al: Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. Proc Natl Acad Sci USA 2004;101:14240–14245.
- 8 Ussery DW, Borini S, Wassenaar TM: Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists (Computational Series). Springer Verlag London, 2008.
- 9 Ong C, Ooi CH, Wang D, Chong H, Ng KC, et al: Patterns of large-scale genomic variation in virulent and avirulent *Burkholderia* species. Genome Res 2004;14:2295–2307.
- 10 Kim HS, Schell MA, Yu Y, Ulrich RL, Sarría SH, et al: Bacterial genome adaptation to niches: divergence of the potential virulence genes in three *Burkholderia* species of different survival strategies. BMC Genomics 2006;6:174.
- 11 Chain PS, Denev VJ, Konstantinidis KT, Vergez LM, Agulló L, et al: *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci USA 2006;103:15280–15287.
- 12 Winsor GL, Khaira B, Rossum TV, Lo R, Whiteside MD, Brinkman FS: The *Burkholderia* Genome Database: facilitating flexible queries and comparative analysis. Bioinformatics 2008;24:2803–2804.
- 13 Fukuchi S, Nishikawa K: Estimation of the number of authentic orphan genes in bacterial genomes. DNA Res 2004;11:219–231.
- 14 Nielsen P, Krogh A: Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics 2005;21:4322–4329.
- 15 Larsen TS, Krogh A: EasyGene – a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics 2003;4:21.
- 16 Binnewies TT, Hallin PF, Staerfeldt HH, Ussery DW: Genome Update: proteome comparisons. Microbiology 2005;151:1–4.
- 17 Hallin PF, Binnewies TT, Ussery DW: The genome BLAST atlas – a GeneWiz extension for visualization of whole-genome homology. Mol Biosyst 2008;4:363–371.
- 18 Holden MT, Seth-Smith HM, Crossman LC, Sebahia M, Bentley SD, et al: The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol 2009;191:261–277.
- 19 Wassenaar TM, Bohlin J, Binnewies TT, Ussery DW: Genome comparison of bacterial pathogens. Genome Dyn 2009;6:1–20.
- 20 Tuanyok A, Leadem BR, Auerbach RK, Beckstrom-Sternberg SM, Beckstrom-Sternberg JS, et al: Genomic islands from five strains of *Burkholderia pseudomallei*. BMC Genomics 2008;9:566.
- 21 Lan R, Reeves PR: Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 2000;8:395–401.
- 22 Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al: Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial ‘pan-genome’. Proc Natl Acad Sci USA 2005;102:13950–13955.
- 23 Sim SH, Yu Y, Lin CH, Karuturi RK, Wuthiekanun V, et al: The core and accessory genomes of *Burkholderia pseudomallei*: implications for human melioidosis. PLoS Pathogens 2008;4:e1000178.
- 24 Tayeb LA, Lefevre M, Passet V, Diancourt L, Brisse S, Grimont PA: Comparative phylogenies of *Burkholderia*, *Ralstonia*, *Comamonas*, *Brevundimonas* and related organisms derived from *rpoB*, *gyrB* and *rrs* gene sequences. Res Microbiol 2008; 159:169–177.
- 25 Godoy D, Randle G, Simpson AJ, Aanensen DM, Pitt TL, et al: Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. J Clin Microbiol 2003;41:2068–2079. Erratum in: J Clin Microbiol 2003;41:4913.
- 26 Baldwin A, Mahenthalingam E, Thickett KM, Honeybourne D, Maiden MC, et al: Multilocus sequence typing scheme that provides both species and strain differentiation for the *Burkholderia cepacia* complex. J Clin Microbiol 2005;43:4665–4673.
- 27 Bohlin J, Skjerve E, Ussery DW: Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. BMC Genomics 2008;9:104.

David W. Ussery
 Center for Biological Sequence Analysis, Department of Systems Biology
 Building 208, Technical University of Denmark
 DK–2800 Lyngby (Denmark)
 Tel. +45 45 25 24 88, Fax +45 45 93 15 85, E-Mail dave@cbs.dtu.dk

© Free Author
 Copy – for personal use only

ANY DISTRIBUTION OF THIS ARTICLE WITHOUT WRITTEN CONSENT FROM S. KARGER AG, BASEL IS A VIOLATION OF THE COPYRIGHT.

Written permission to distribute the PDF will be granted against payment of a permission fee, which is based on the number of accesses required. Please contact permission@karger.ch