The Core and Accessory Genomes of Burkholderia pseudomallei: Implications for Human Melioidosis

Siew Hoon Sim1*, Yiting Yu2*, Chi Ho Lin2, R. Krishna M. Karuturi2, Vanaporn Wuthiekanun3, Apichai Tuanyok4, Hui Hoon Chua2, Catherine Ong1, Sivalingam Suppiah Paramalingam1, Gladys Tan1, Lynn Tang1, Gary Lau1, Eng Eong Ooi1, Donald Woods4, Edward Feil5, Sharon J. Peacock3,6*, Patrick Tan2,7*

1 Defense Medical and Environmental Research Institute, DSO National Laboratories, Singapore, Republic of Singapore, 2 Genome Institute of Singapore, Singapore, Republic of Singapore, 3 Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand, 4 Faculty of Medicine, University of Calgary Health Sciences Centre, Calgary, Alberta, Canada, 5 Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, United Kingdom, 6 Center for Clinical Vaccinology and Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, Headington, Oxford, United Kingdom, 7 Duke-NUS Graduate Medical School Singapore, Singapore, Republic of Singapore

Abstract

Natural isolates of Burkholderia pseudomallei (Bp), the causative agent of melioidosis, can exhibit significant ecological flexibility that is likely reflective of a dynamic genome. Using whole-genome Bp microarrays, we examined patterns of gene presence and absence across 94 South East Asian strains isolated from a variety of clinical, environmental, or animal sources. 86% of the Bp K96243 reference genome was common to all the strains representing the Bp “core genome”, comprising genes largely involved in essential functions (eg amino acid metabolism, protein translation). In contrast, 14% of the K96243 genome was variably present across the isolates. This Bp accessory genome encompassed multiple genomic islands (GIs), paralogous genes, and insertions/deletions, including three distinct lipopolysaccharide (LPS)-related gene clusters. Strikingly, strains recovered from cases of human melioidosis clustered on a tree based on accessory gene content, and were significantly more likely to harbor certain GIs compared to animal and environmental isolates. Consistent with the inference that the GIs may contribute to pathogenesis, experimental mutation of BPSS2053, a GI gene, reduced microbial adherence to human epithelial cells. Our results suggest that the Bp accessory genome is likely to play an important role in microbial adaptation and virulence.

Introduction

Melioidosis is a potentially fatal infectious disease of humans and animals caused by the Gram-negative bacterium Burkholderia pseudomallei (Bp) [1]. An environmental saprophyte found in South East Asia, Bp infections in endemic areas may be responsible for up to 20% of deaths due to septicemia [2,3], and Bp has been designated a Category B bioterrorism agent [4]. A wide spectrum of disease symptoms are associated with melioidosis often leading to late diagnosis and treatment [5]. Commonly presenting as an acute septicemic illness, chronic Bp infection is also well recognized which can be confused with TB or malignancy [6]. Besides humans, Bp has a broad host range and can infect nematodes, amoebae, dolphins, birds, swine, sheep, and gorillas [7–11]. Bp can also be isolated from diverse environmental sources such as soil, water, and air [12–17]. Identifying the molecular factors responsible for this tremendous ecologic flexibility may improve our understanding of microbial survival and adaptation, and suggest novel diagnostic and treatment strategies for melioidosis.

The phenotypic versatility of Bp is likely to be underpinned by the presence of a highly dynamic genome. For example, lateral gene transfer events may cause large-scale variations in genome content [18]. The portion of the genome that is variably present between individual strains is often termed the “accessory genome”, to distinguish these genes from genes common to all strains in a population and involved in essential functions (the “core” genome). In several microbial species, accessory genes have been shown to play key roles in host adaptation and, in the case of Bp, the accessory genome may contribute to virulence and antibiotic resistance [19]. Interestingly, previous studies indicate that in Bp, gene loss, as well as gene acquisition events, can both cause phenotypic shifts towards virulence. For example, comparisons between Bp and B. thailandensis, an avirulent closely related species, have shown that an important evolutionary step in the development of Bp pathogenicity was the loss of an anti-virulence arabinose assimilation cluster [20,21]. Such findings thus raise a compelling need to accurately define the core and accessory genomes of Bp.

In other γ proteobacteria genera (E. coli, Pseudomonas, Vibrio), the accessory genome can encompass up to 20% of all genomic content, and similar percentages may also hold for Burkholderia spp. [22–23]. However, to date, comprehensive qualitative and quantitative studies of the core and accessory genome in Bp have not been carried out, and the full extent to which gene content differences contribute to virulence in Bp is still unclear. While...
Author Summary

Melioidosis is a serious infectious disease of humans caused by *Burkholderia pseudomallei*, a soil bacterium endemic to many areas in South East Asia. Besides humans, *B. pseudomallei* is also capable of infecting many other species and can be isolated from diverse environmental sources including soil, water, and air. In this study, we used DNA microarrays to probe the stability of the *B. pseudomallei* genome in a large panel of clinical, animal, and environmental strains. We found that evidence of a highly dynamic *B. pseudomallei* genome, with up to 14% being variably present across different strains. Surprisingly, strains recovered from human patients were significantly associated with the presence of “genomic islands”, corresponding to regions of DNA directly acquired from other microorganisms. Genes on these genomic islands may thus play an important role in the pathogenesis of human melioidosis.

some previous studies have attempted to explore these issues, they have not incorporated data from the whole genome [19,26–28], or have used only a very small sample of strains [29,30]. In this study, we performed a detailed array-based comparative genomic hybridization (aCGH) analysis of close to 100 clinical, animal and environmental Bp isolates from South East Asia. To our knowledge, this is the first time a whole genome comparative study has been applied to such a large Bp strain cohort. We found that 86% of the reference Bp K96243 genome was present in all the strains, while the remaining 14% was variably present across the strain panel. Surprisingly, isolates associated with human melioidosis exhibited a tendency to harbor certain GIs compared to strain panel. Surprisingly, isolates associated with human melioidosis may thus play an important role in the pathogenesis of human melioidosis.

**Results**

Genome-wide Identification of Core and Accessory Genes in Bp Isolates

Using a previously validated Bp K96243 DNA microarray [30,31], we generated aCGH profiles for ninety-four Bp strains isolated from human patients, animals, and environmental soils in Singapore, Malaysia or Thailand (Table S1). We applied a Gaussian Mixture Model (GMM) to the aCGH data and identified 750 out of 5369 genes (14%) as being variably present across the strain panel (see Methods and Figure S1). The variability of the 750 genes was experimentally validated by several independent methods, including bioinformatic comparisons to previously-known variable genes, comparisons against publicly available genome sequences, and experimental confirmation by targeted PCR assays (Figure S2 and Table S2).

The Bp Core Genome Encodes Essential Processes and a Common Virulence Machinery

86% of the Bp K96243 genes (4619) were found in all strains, representing the Bp core genome (Figure 1). Using pathway analysis, we found that the core genes were significantly over-represented in several functions necessary for basic bacterial growth and survival, including amino acid metabolism ($1.52 \times 10^{-5}$), inorganic ion transport ($3.96 \times 10^{-5}$), nucleotide metabolism ($1.52 \times 10^{-5}$) and protein translation ($7 \times 10^{-5}$) (Table 1). The core genes were also significantly enriched in genes conserved in other *Burkholderia* species (Bp, B. mallei, B. thailandensis and *B. cepacia*) ($p = 8.68 \times 10^{-11}$) (Text S1 and Table S3), suggesting that a significant proportion of these Bp core genes may represent core genes in other related species as well [32].

Besides these basic housekeeping functions, the Bp core genes were also significantly enriched in commonly encountered virulence-related genes such as secretion proteins, capsular polysaccharides, exoproteins, adhesins, fimbriae and pili ($p = 1.8 \times 10^{-5}$) (Table 1).

For example, three Bp-specific fimbral gene clusters (BPSL1626-1629, BPSL1799-1801, BPSS0120-0123) were found in all strains. This finding suggests that most, if not all, Bp isolates are likely to possess a common ‘virulence machinery’. Notably, many of these conventional virulence genes are also found in other related species such as *B. thailandensis* that although non-infectious to

![Figure 1. The Core and Accessory Genomes of Bp.](doi:10.1371/journal.ppat.1000178.g001)
mammals can kill other species such as nematodes [20,33]. This is consistent with the possibility that Bp might have descended from a pathogenic ancestor with a non-mammalian host.

Functional and Chromosomal Biases in the Bp Accessory Genome

14% of the Bp K96243 genome was variable across the strain panel, representing the Bp accessory genome. Since our analysis is confined to genetic elements present in the reference K96243 genome, the extent of genomic variability reported here should be regarded as a lower limit. The 750 variable genes were equally distributed between both Chromosome 1 and Chromosome 2 after normalizing for chromosome size differences. The accessory genes were significantly enriched in paralogous genes (p = 2 × 10^{-7}) and genes encoding hypothetical proteins (p = 3 × 10^{-6}) (Table 1). Approximately one-third (30.8%) of the accessory genes were localized to a series of previously identified “genomic islands” (GIs) in the K96243 genome [34]. GIs are regions bearing unusual sequence hallmarks, such as atypical GC content and/or dinucleotide frequencies, and are likely to have been recently acquired by lateral gene transfer. Of sixteen GIs in the K96243 genome, fourteen GIs were represented by accessory genes. In contrast, two GIs (7 and 14) were found in all strains, suggesting that GIs 7 and 14 should be regarded as part of the Bp core genome.

Besides the GIs, we also identified several novel regions of at least three contiguous probes that were absent in at least three strains. Henceforth referring to these regions as ‘indels’, we identified eight indels on chromosome 1, and twelve on chromosome 2 (Table 2). We experimentally validated two of these indels using PCR assays (Figure S3). The indels ranged in size from 1.3 to 7.5 kb, and were absent in 12.9% to 45.2% of strains (Figure 2). Three indels (n1, n4 and n11) were associated with atypical GC content (53.7–38.6%, compared to 68% for the Bp genome), and four (n2, n9, n11 and n16) carried genes characteristic of mobile genetic elements such as integrases, transposases and bacteriophage-related genes, consistent with lateral transfer. These indels may therefore share similar dynamics to the larger genomic islands, and may be considered as genomic “islets”. In other species, analogous islets which are typically <10 kb long, have been shown to play a role in virulence (e.g. the sifA islet in S. typhimurium) [35]. Of note, n16 and n18 were flanked at both their 5’ and 3’ ends by tandem repeat sequences, while n4, n6, n8 and n19 possessed sequence repeats at either their 5’ or 3’ ends. In some cases, the indels in the Bp genomes may actually form part of the larger GIs. For example, n2 (BPSL0741-BPSL0744) was located at the 5′ boundary of GI 4 (BPSL0743-BPSL0772), while n11 (BPS0393-BPS0397) was located immediately 3′ to GI 15 (BPS0378-BPS0391A).

Three indel regions (n6, n12 and n19) contained genes associated with LPS metabolism. Lipopolysaccharides (LPS) are macromolecular components on the outer membranes of Gram-negative bacteria composed of lipid A, core oligosaccharide, and O-antigen polysaccharides [36]. LPS molecules are commonly immunogenic and have been previously implicated in virulence for numerous microbes [37,38]. Region n6 (BPSL2666-BPSL2668) contains a phosphoglucomutase (BPSL2666), a lipopolysaccharide LPS biosynthesis protein (BPSL2667) and a glycosyltransferase (BPSL2668), and was located four genes away from a larger LPS biosynthesis cluster (BPSL2672-BPSL2680). Both regions n12 (BPS0427 - BPS0429) and n19 (BPS2245-BPS2255) contained two O-antigen related genes, including O-acetyltransferase and glycosyltransferase. While n12 corresponds to a previously identified type III O-PS polysaccharide gene cluster [39], the contribution of n19 genes to Bp LPS biology is currently unknown. The identification of three physically unlinked indels related to LPS metabolism provides a mechanism by which high levels of LPS diversity may be maintained in the Bp population [40].

Unsupervised Clustering Using the Accessory Genome Distinguishes Clinical Isolates from Animal and Environmental Strains

To explore if differences in accessory genome content might be associated with host adaptation or the propensity to cause disease, we applied unsupervised clustering to cluster the strains using the entire set of 750 accessory genes (“accessory genome clustering”, AGC). We identified three large AGC clusters each containing 27 to 42 strains, with each cluster containing at least 4–6 sub-branches (Figure 3). Most strikingly, the majority of human clinical isolates (73.1%) fell into one AGC cluster (Clade C), another cluster contained 73.7% of the animal isolates (Clade A), and a third cluster contained 45% of the environmental isolates (Clade E). Similar results were obtained when the clustering was repeated using either Chromosome 1 or Chromosome 2 accessory genes (Figure S4). The over-representation of human clinical isolates in the C clade was highly significant (P = 2.001 × 10^{-14}, Fisher’s exact test), and of the remaining 13 clinical isolates nine segregated within the E clade and four in the A clade. This clustering pattern is unlikely to represent differences in geographical distribution, since the majority of the clinical (65%), animal (89%) and environmental isolates (80%) were isolated in Singapore within a ~700 km² region or from nearby islands. Furthermore, clinical isolates from Thailand clustered with the other clinical isolates, despite being geographically remote. This analysis therefore suggests that strains associated with human melioidosis may possess an accessory genome distinct from most animal and environmental strains. We also note that all three clades contained environmental isolates, which is consistent with the view that the environment represents a diverse reservoir from which human and animal adapted strains emerge.

**Table 1. Enriched Functions of Core and Accessory Genes in Bp.**

<table>
<thead>
<tr>
<th>Gene Distribution</th>
<th>Accessory</th>
<th>Core</th>
<th>Total</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Genes</td>
<td>750</td>
<td>4619</td>
<td>5369</td>
<td></td>
</tr>
<tr>
<td><strong>Enriched Functions in Core Genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid transport and metabolism*</td>
<td>37</td>
<td>377</td>
<td>414</td>
<td>1.5 × 10^{-3}</td>
</tr>
<tr>
<td>Inorganic ion transport and metabolism*</td>
<td>16</td>
<td>199</td>
<td>215</td>
<td>3.96 × 10^{-3}</td>
</tr>
<tr>
<td>Nucleotide transport and metabolism*</td>
<td>4</td>
<td>78</td>
<td>82</td>
<td>0.0152</td>
</tr>
<tr>
<td>Protein Translation*</td>
<td>12</td>
<td>158</td>
<td>170</td>
<td>0.007</td>
</tr>
<tr>
<td>Virulence Components*</td>
<td>30</td>
<td>321</td>
<td>351</td>
<td>1.83 × 10^{-3}</td>
</tr>
<tr>
<td><strong>Enriched Functions in Accessory Genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paralogous Genes</td>
<td>73</td>
<td>228</td>
<td>301</td>
<td>2.25 × 10^{-7}</td>
</tr>
<tr>
<td>Hypothetical Proteins</td>
<td>233</td>
<td>1132</td>
<td>1365</td>
<td>3.3 × 10^{-4}</td>
</tr>
</tbody>
</table>

P-values were computed using a Fisher Test.

*P-values were computed based upon the simultaneous comparison of 25 COG pathways.

Virulence genes were obtained from an annotated listing provided in Holden et al (2004) [34].

doi:10.1371/journal.ppat.1000178.t001

*P-values were computed based upon the simultaneous comparison of 25 COG pathways.

**Table 1. Enriched Functions of Core and Accessory Genes in Bp.**

<table>
<thead>
<tr>
<th>Gene Distribution</th>
<th>Accessory</th>
<th>Core</th>
<th>Total</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Genes</td>
<td>750</td>
<td>4619</td>
<td>5369</td>
<td></td>
</tr>
<tr>
<td><strong>Enriched Functions in Core Genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid transport and metabolism*</td>
<td>37</td>
<td>377</td>
<td>414</td>
<td>1.5 × 10^{-3}</td>
</tr>
<tr>
<td>Inorganic ion transport and metabolism*</td>
<td>16</td>
<td>199</td>
<td>215</td>
<td>3.96 × 10^{-3}</td>
</tr>
<tr>
<td>Nucleotide transport and metabolism*</td>
<td>4</td>
<td>78</td>
<td>82</td>
<td>0.0152</td>
</tr>
<tr>
<td>Protein Translation*</td>
<td>12</td>
<td>158</td>
<td>170</td>
<td>0.007</td>
</tr>
<tr>
<td>Virulence Components*</td>
<td>30</td>
<td>321</td>
<td>351</td>
<td>1.83 × 10^{-3}</td>
</tr>
<tr>
<td><strong>Enriched Functions in Accessory Genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paralogous Genes</td>
<td>73</td>
<td>228</td>
<td>301</td>
<td>2.25 × 10^{-7}</td>
</tr>
<tr>
<td>Hypothetical Proteins</td>
<td>233</td>
<td>1132</td>
<td>1365</td>
<td>3.3 × 10^{-4}</td>
</tr>
</tbody>
</table>

P-values were computed using a Fisher Test.

*P-values were computed based upon the simultaneous comparison of 25 COG pathways.

Virulence genes were obtained from an annotated listing provided in Holden et al (2004) [34].

doi:10.1371/journal.ppat.1000178.t001
Clinical Isolates are Associated with the Presence of Genomic Islands

We then performed a supervised analysis to identify which of the 750 accessory genes were significantly different between the C and A/E clades. Of the 750 genes, 218 genes were commonly present in isolates in the C clade but absent from strains in the other two clusters (Figure 4A). Strikingly, we found that almost all of these 218 genes (85%) were localized to the GIs, with all fourteen GIs being represented. This figure (85%) is significantly higher than the 31% of all accessory genes located on GIs, raising the possibility that GIs may play an important role in determining ecological niche and host adaptation.

Is there any direct evidence that genes encoded on GIs, and which define the C clade, might play an important role in the biology or pathogenicity of Bp? Unfortunately, almost 35% of the GI genes encode ‘hypothetical’ proteins (Table S4), meaning that their function is unknown. For those genes specific to the C clade where functions could be assigned, several broad functional classes were represented. For example, GB8 contains several genes spermidine/putrescine transport genes ($potB$, $potC$, $potG$), which have been associated with biofilm formation and the regulation of...
Type III secretion genes [41,42]. Type I restriction-modification enzymes are found on GI5 and GI10, and a glutathione S-transferase gene (BPSS2048) on GI16 may impart resistance to oxidative stress. Also supporting their potential role in Bp biology, several GI genes exhibited distinct and complex gene expression patterns during Bp growth (Text S2). However, the role of such genes in pathogenesis remains speculative. In order to explore this further, we generated an experimentally mutated strain (ATS2053) disrupted in BPSS2053, a GI 16 gene encoding a hemagglutinin-related protein, and determined the adherence of the mutant strain to human buccal epithelial cells. A highly significant reduction in the adherence to buccal epithelial cells was noted between the 1026b clinical isolate and the isogenic ATS2053 mutant strain (mean adherence: 1026b - 16.3 ± 3.2 vs ATS 2053 - 4.4 ± 1.7, p < 0.001, Students t test). This finding provides evidence pointing both to the biological relevance of GI genes, but more specifically to a role of these genes in virulence.

Comparison of Clustering Patterns Using MLST or the Accessory Genome

Finally, we examined the concordance between strain clusters defined on the basis of accessory gene content and the phylogenetic signal within the Bp core genome. We characterised 45 representative isolates by Multilocus Sequence Typing (MLST), a typing scheme that indexes variation at seven core housekeeping genes [43]. Using the previously published Bp scheme [44], we resolved the 45 isolates into 9 sequence types (ST 46, 51, 54, 84, 169, 289, 414, 422 and 423). Seven of these STs (ST51, 54, 84, 46, 169, 289, 414) have been previously observed in Malaysia, Thailand, and Singapore [44,45].

Previous analyses of MLST for Bp have highlighted the difficulties in building robust phylogenetic trees for this species, owing to a paucity of informative sites in the concatenated data and frequent homologous recombination [46]. We thus favored a categorical approach to comparing the AGC and MLST data by examining the distribution of sequence types across the three clades defined by the AGC data (Table 3). This analysis revealed that the STs are not randomly distributed between the three clusters, indicating some consistency between the MLST and AGC datasets. Most strikingly, of the 20 ST51 isolates, 17 clustered within the animal-associated clade (A), three within the clinical C clade, and none in the environmental E clade. Of the other STs where at least 4 isolates were observed, all four ST422 isolates corresponded to the C clade, and all four ST84 isolates clustered within the E clade. Finally, of the nine ST423 isolates, five clustered within the C clade and four in the E clade. These data suggest that the animal-associated clade is likely to correspond to a single clone (ST51) and provides some evidence for concordance between STs 422 and 84 with the AGC data, although the evidence in these latter cases is equivocal due to the small number of strains. In contrast, the “split” of the ST423 isolates between the clinical and environmental clades, and the 3 ST51 isolates belonging to the clinical clade, represent clear discrepancies between the two datasets. Possible explanations for these discrepancies may represent convergence of either the MLST or the AGC data, as discussed below.

Discussion

In this report, we present a comprehensive aCGH analysis for a large series of natural Bp isolates. We found that the accessory (variably present) portion of the Bp genome corresponds to ~14% of the whole genome content, which is broadly similar to other γ-proteobacteria. Since this approach is limited to the detection of elements present in the Bp K96243 genome, and novel elements in query genomes are not detected, this estimated fraction of the accessory genome should be regarded as a lower bound. In the only published study of a Bp genome sequence to date, Holden et al (2004) computationally identified 16 GIs comprising 6% of the K96243 genome [34], and our data confirm that most of these islands are indeed highly variable between strains.
However, two GIs (7 and 14) were found in all strains and should thus be regarded as part of the Bp core genome. Furthermore, our data also revealed the variable presence of several other small genomic islets/indels across the two chromosomes, which might contribute to the phenotypic diversity of Bp. Notably, we observed that several indels (n6, n12 and n19) were related to LPS biology. Currently, the exact contribution of LPS to Bp virulence is unclear. For example, DeShazer et al (1998) showed that Bp type II O-PS is essential for serum resistance and virulence [47], and mice pre-immunized with Bp LPS displayed enhanced survival to a subsequent challenge [48]. In contrast, other groups have reported that Bp LPS exhibits a reduced ability to activate immune cells compared to E. coli LPS, suggesting that LPS might play only a minimal role in Bp virulence. It is possible that these conflicting results might reflect heterogeneity in LPS pathways resulting from the variable presence of these indels, and represent an important mechanism for host adaptation. Interestingly, while it was recently shown that type III O-PS mutants (indel n12) do not appear to exhibit significant virulence attenuation in mouse infection assays [39], we have found in preliminary work that Bp strains lacking the indel n19 LPS cluster generally exhibited lower levels of virulence compared to strains where this cluster was present (SSH, data not shown). In the AGC tree, n19 was absent both from three strains segregating as a single branch in the A clade, and from 5

Figure 3. Unsupervised Accessory Genome Clustering of Bp Isolates. Clustering diagram of Bp strains on the basis of accessory genome content. The tree is constructed using MultiExperiment Viewer (MeV) version 4, based on the entire 750-gene accessory genome and combined average linkage hierarchical clustering. Clinical (labeled in red), Animal (labeled in blue) and Environmental (labeled in green) strains are indicated. Isolates from Thailand are highlighted in the red broken circle. Three broad clusters/clades are identified which are named C-clinical, A-animal, E-environmental, with the percentage of concordant strains in that cluster. Numbers on branches represent bootstrap values based on 1000 tests. The bootstrapping analysis reveals a clear distinction between the C (clinical) and A/E clusters (non-clinical - animal and environmental) (Bootstrap value = 100).

doi:10.1371/journal.ppat.1000178.g003
The role of these LPS clusters in Bp virulence. Further experiments are clearly required to understand the significance of these findings. It thus seems entirely plausible that genes facilitating the production of bacteriocides from other bacteria and phage prophages might this biological role be? At present, we consider it most likely that this scenario is, on balance, less likely. Another possibility that might explain the enrichment of GIs in the clinical isolates is that Bp is undergoing cryptic cycling through the normal human host population could also lead to the selection of factors that promote survival in vivo. However, as we consider the human host to be a relatively minor component of Bp ecology, we argue that this scenario is, on balance, less likely. The availability of both MLST and aCGH data for a representative sub-sample of isolates also provided us the opportunity to compare clade distributions defined either by accessory genome content or allelic variation in the core genome. We found that the animal associated strains largely corresponded to a single MLST clone (ST51). These isolates were assembled from three distinct sources: the Singapore zoo, the University of Malaya and a pig abattoir in Singapore. The soil isolates corresponding to ST51 (which also clustered in the A clade) were not isolated from soil samples in proximity to the animal ST51 isolates, which suggests that this genotype is also present in the environment. The homogeneity of these isolates is therefore striking and cannot be explained simply by sampling bias. The consistency between the microarray and MLST data strongly suggest that this clade is monophyletic, and that the strains harbour similar gene repertoires by virtue of common descent. In contrast, we also observed clear discrepancies between the MLST and aCGH clades. For example, three ST51 isolates

strains in the C clade that segregated across multiple branches. This suggests that n19 may have been recurrently lost in different Bp lineages. Further experiments are clearly required to understand the role of these LPS clusters in Bp virulence.

We also found that the Bp strains could be clustered into distinct clades based on both the presence and absence of specific accessory genes. Of primary interest, strains belonging to the C clade of clinical isolates were largely defined by the presence of 218 genes, of which 85% are localized to the GIs. These findings provide evidence for a distinct repertoire of Bp genes that may cause a predisposition to human disease and that these genes tend to be located on GIs. Although many of the genes encoded on the GIs are of unknown function, we present experimental evidence that a strain mutated in one of these genes exhibited decreased adherence to human buccal endothelial cells, supporting a role in virulence potential. We also observed coordinated growth-associated expression of several GI genes, which is also consistent with the view that they play an important biological role. What might this biological role be? At present, we consider it most likely that this “virulent” combination of genes has likely emerged for reasons other than to cause human disease, particularly since cases of human (or animal) infection are relatively rare compared to the density of Bp in the soil. In contrast to bacteria which are obligately associated with eukaryotic hosts, soil bacteria such as Bp commonly face extreme and unpredictable biotic and abiotic challenges including extreme temperature shifts, solar radiation, variable humidity, competition for nutrients, and the requirement to survive ingestion by predatory protozoa, nematodes, the production of bacteriocides from other bacteria and phage infection. It thus seems entirely plausible that genes facilitating survival against these environmental challenges might have also indirectly enhanced the microbe’s ability to colonize and “accidently” infect a human host, particularly when the host is immunocompromised. Another possibility that might explain the enrichment of GIs in the clinical isolates is that Bp is undergoing cryptic cycling through the normal human host population could also lead to the selection of factors that promote survival in vivo. However, as we consider the human host to be a relatively minor component of Bp ecology, we argue that this scenario is, on balance, less likely. The availability of both MLST and aCGH data for a representative sub-sample of isolates also provided us the opportunity to compare clade distributions defined either by accessory genome content or allelic variation in the core genome. We found that the animal associated strains largely corresponded to a single MLST clone (ST51). These isolates were assembled from three distinct sources: the Singapore zoo, the University of Malaya and a pig abattoir in Singapore. The soil isolates corresponding to ST51 (which also clustered in the A clade) were not isolated from soil samples in proximity to the animal ST51 isolates, which suggests that this genotype is also present in the environment. The homogeneity of these isolates is therefore striking and cannot be explained simply by sampling bias. The consistency between the microarray and MLST data strongly suggest that this clade is monophyletic, and that the strains harbour similar gene repertoires by virtue of common descent. In contrast, we also observed clear discrepancies between the MLST and aCGH clades. For example, three ST51 isolates

<table>
<thead>
<tr>
<th>AGC Clades</th>
<th>C</th>
<th>A</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLST Sequence Types</td>
<td>ST51</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>ST423</td>
<td>5</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>ST422</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>ST84</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>ST169</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ST46</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ST54</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ST414</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ST289</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>17</td>
<td>13</td>
</tr>
</tbody>
</table>

Depicted are the distributions of 45 Bp strains subjected to both AGC and MLST analysis. Strain numbers in bold (eg ST51) highlight STs where the majority of strains were found in one AGC clade.

doi:10.1371/journal.ppat.1000178.003
clustered within the clinical aCGH clade, and ST423 was split between the clinical and environmental aCGH clades. There are three possibilities to explain these discrepancies: i) The MLST data represents the ancestral state which is inherited by descent into two AGC-defined clades - this is unlikely for the animal cluster as the vast majority of isolates are ST31, but might conceivably explain the ST423 split between the clinical and environmental clades. ii) Convergence of the MLST alleles - this would imply that isolates with the same ST are not identical by descent but happen to share the same combination of alleles. The presence of a very common alleles for each gene, combined with high rates of recombination in Bp make this possibility more likely. iii) Independent convergence of gene content to one of the three clusters. Unless large numbers of genes can be transferred in single events, this possibility seems less parsimonious than (ii). More data are required to examine which of these hypotheses is most likely.

In summary, our study provides direct experimental confirmation that the Bp genome is highly plastic, and that gene acquisition and deletion are major drivers of this variability. This variability is far from random, and is functionally biased towards genes involved in mobile elements, hypothetical and paralogous genes, and LPS biosynthesis. Furthermore, genes on mobile elements may predispose individual strains, either directly or indirectly, towards causing human disease. We believe this latter result is significant in that most Bp research to date has focused on virulence components in the Bp core genome rather than genes on mobile elements. We conclude by noting that most of the Bp genome sequences currently available have been obtained from human clinical isolates. Given our results, it might be highly informative to subject a panel of animal and environmental Bp isolates to similar detailed genome analysis as well.

Methods

Bacterial Strains

Ninety-four Bp isolates were used in this study. These include: a) the K96243 reference strain, b) 32 clinical isolates from melioidosis patients between 1996 and 2005, c) 19 animal isolates from various species (eg monkeys, pigs, birds, and dogs) diagnosed with melioidosis between 1996 and 2000, d) 20 soil isolates from 1994 to 2003, and e) two type strains (ATCC23343 and ATCC15682). All strains were isolated in Singapore, neighboring islands, or surrounding countries (Malaysia, Thailand). The isolates were sampled from a diversity of locations and not a single site, supporting their unbiased nature (As Lay Tin and Joseph Tong, personal communication). Further strain information is provided in Table S1.

Genomic DNA Extraction and Array-Based Comparative Genomic Hybridization (aCGH)

Strains were cultured on Tryptone Soy Agar (TSA) (Difco Laboratories, Detroit, Michigan) at 37°C, and genomic DNA extracted using a genomic DNA purification kit (Qiagen). The Bp DNA microarray has been previously described [29-31] and comprises approximately 16,000 PCR-amplified array probes representing all 5742 predicted genes in the K96243 genome printed in duplicate. Test genomic DNA (2 μg) was fluorescently labeled with Cy3-dCTP (Amersham Pharmacia Biotech) using nick-translation and co-hybridized to the array with an equal quantity of Cy5-dCTP (Amersham Pharmacia Biotech) labeled reference K96243 DNA. The absence of significant dye-bias artifacts was confirmed by analyzing reciprocal dye-swap hybridizations for 10 isolates data not shown, also see ref [29]. Raw fluorescence data was acquired using an Axon scanner with GENEPIX v4.0 software (Axon Instruments, Redwood City, CA).

Microarray Data Preprocessing

Individual arrays were internally normalized between the Cy3 and Cy5 channels by LOWESS normalization, and the entire dataset was cross-normalized by median-scaling each array to the same Cy3/Cy5 ratio. To filter the microarray data, we eliminated probes exhibiting a missing value score across >40% of samples (indicating that they were not reliably measured), and probes whose genomic loci were redundant with other probes. This data filtering procedure generated a final high-quality data set of 5369 non-redundant probes. The entire microarray data set is available at the Gene Expression Omnibus database under accession number GSE9491.

Identification of Accessory Genes

A Gaussian mixture model (GMM) [51] was used to identify accessory and core genes in the data set. In concept, a GMM fits a test signal distribution (such as microarray data) to either a single or double gaussian curve, and the likelihood that the distribution corresponds to a single curve is computed. The GMM was applied in two stages. First, p-values were computed using the aCGH profiles of each individual array spot, following a chi-square distribution with 3 degrees of freedom under the null hypothesis that the data distribution of the spot follows a 1-gaussian distribution. Second, since each probe was spotted twice on the array, we obtained composite p-values of each array probe using Inverse Chi-square Meta-Analysis [32], squaring the p-values of both spots belonging to the same probe. This latter statistic follows a chi-square distribution with 4 degrees of freedom. All p-values were corrected for multiple-hypothesis testing according to the Benjamini-Hochberg procedure [53]. A cut-off of p=1.83E-08 was selected to define the top 750 most highly variable probes, representing the accessory genome.

Pathway Analysis of Core and Accessory genes

All protein coding sequences in the Bp K96243 genome were queried by BLASTP against the Cluster of Orthologous group (COGs) database, a public bioinformatic database that groups protein sequences on the basis of phylogenetic similarity to various cellular functions, such as protein translation, DNA replication and transcription, nuclear structure and defense mechanisms (accessible at http://www.ncbi.nlm.nih.gov/COG/new/). Matches were defined as database hits with an e-value threshold of <10^-10. Based on the COG assignments, the K96243 proteins were assigned to functional categories. Fisher’s exact tests were used to identify significantly overrepresented COG categories in either the core or accessory genes. To identify conserved genes (metagenes) across four Burkholderia species, we queried the 3460 Chr 1 and 2395 Chr 2 ORFs in the Bp K96243 genome against the B. cenocepacia (Bc), B. mallei (Bm), and B. thailandensis (Bt) genomes using tblastn [32] (Text S1). To minimize the number of ambiguous predictions including ORFs with matches to multiple genomic locations, we constrained the resulting matches to have I) a minimum length of 50 amino acids, II) a minimal e-value cutoff of 1e-6 and III) a minimum percent identity of 50%. Homology assignments returned 2673 genes and were validated by a reciprocal blast assay resulting in 2590 genes. Control analyses using either Bc, Bm or Bt as starting reference genomes yielded similar metagene sets (data not shown). Paralogous genes were identified using the CD-HIT program [54] as genes with >60% identity to one another, following established studies [53,56]. Tandem repeat regions in the K96243 genome were identified using the Tandem Repeats Finder program [57].

Clustering Analysis

Phylogenetic trees based on aCGH profiles were constructed using MultiExperiment Viewer (MeV) version 4 (http://www.tm4. The Core and Accessory Genomes of B. pseudomallei
Construction of Mutants

The BPSS2053 (fhaB) gene was disrupted in strain DD503, an isogenic derivative of wild-type 1026b. In DD503, the aminoglycoside resistance-determining region from B. pseudomallei 1026b chromosomal DNA as previously described [21]. The PCR product was cloned into pCR2.1-TOPO (Invitrogen Life Technologies, Burlington, Ontario, Canada) according to the manufacturer’s instructions. The internal region from BPSS2053 was cloned as an EcoR1 fragment into pGSV3-lux, a suicide vector containing a promotorless lux operon as a reporter, to create pATS2053. The recombinant plasmid pATS2053 was transformed into E. coli SM10 parsley [62]. Transformed E. coli containing pATS2053 were conjugated with Bp DD503, and transconjugants selected on LB-gentamicin-polymyxin B agar. The transconjugants were screened for lux-mediated light production by assaying 100 colonies, using the tree files visualized using Phylodraw [60].

MLST allele database.

The Core and Accessory Genomes of B. pseudomallei

MLST on 45 strains was performed as described in Godoy et al (2005) [44] using primer pairs for seven housekeeping genes (ace, glnG, glfB, gmbD, lepA, lipA, narK ndh) on Bp chromosome 1. A complete list of primer pairs and PCR conditions is provided in Table S5. Alleles at each of the MLST loci were assigned using the B. pseudomallei MLST website (http://bpseudomallei.mlst.net/) - each allele was assigned a different allele number and the allelic profile (string of seven integers) was used to define the sequence type (ST). Sequences that were not in the database were checked by re-sequencing, assigned as new alleles and deposited in the MLST allele database.

Supporting Information

Figure S1 Gaussian Distribution curves of genes above and below the GMM threshold Found at: doi:10.1371/journal.ppat.1000178.s001 (0.16 MB DOC)

Figure S2 Experimental and Computational Validation of Variable and Stable Genes in Bp

Figure S3 Validation of 2 novel indel regions (n5 and n7) using PCR

Figure S4 AGC Clusters Based on Chromosome 1 or Chromosome 2

Table S1 Bp Isolates Used in this Study

Table S2 Sequence identities between Bp strains (S13, BP 1710a, 1710b, 1655, Pasteur) Found at: doi:10.1371/journal.ppat.1000178.s006 (0.03 MB DOC)

Table S3 Analysis of conserved metagenes in the set of variable genes and non-variable genes in the B. pseudomallei genome, in the presence and absence of the GI genes

Table S4 Genes Present in Strains Associated with the AGC Clinical Clade

Table S5 Primer pairs for amplification of housekeeping loci in multilocus sequence typing analysis (MLST)

Text S1 Supplementary Methods

Text S2 Expression Patterns of GI Genes During Bp Growth

Acknowledgments

We thank Mongkol Vesaratrachavit and Sarinna Tumapa for their technical assistance.

Author Contributions

Conceived and designed the experiments: SHS YY DW EF PT. Performed the experiments: SHS VW AT HHC CO SSP GT LT GL. Analyzed the data: SHS YY CHL RKKM VW DW EF SJP PT. Contributed reagents/materials/analysis tools: SHS YY CHL RKKM EEO DW EF SJP PT. Wrote the paper: SHS YY CHL RKKM EEO DW EF SJP PT.

References