Mol. Cells *27*, 237-241, February 28, 2009
DOI/10.1007/s10059-009-0029-8

Molecules and Cells

Simple Sequence Repeat (SSR)-Based Gene Diversity in *Burkholderia pseudomallei* and *Burkholderia mallei*

Han Song, Junghyun Hwang, Jaehee Myung, Hyoseok Seo, Hyojeong Yi, Hee-Sun Sim, Bong-Su Kim[1], William C. Nierman[2,3], and Heenam Stanley Kim*

Pathogens *Burkholderia pseudomallei* (Bp) and *Burkholderia mallei* (Bm) contain a large number (> 12,000) of Simple Sequence Repeats (SSRs). To study the extent to which these features have contributed to the diversification of genes, we have conducted comparative studies with nineteen genomes of these bacteria. We found 210 genes with characteristic types of SSR variations. SSRs with nonamer repeat units were the most abundant, followed by hexamers and trimers. Amino acids with smaller and nonpolar R-groups are preferred to be encoded by the variant SSRs, perhaps due to their minimal impacts to protein functionality. A majority of these genes appears to code for surface or secreted proteins that may directly interact with the host factors during pathogenesis or other environmental factors. There also are others that encode diverse functions in the cytoplasm, and this protein variability may reflect an extensive involvement of phase variation in survival and adaptation of these pathogens.

INTRODUCTION

Burkholderia pseudomallei (Bp) and *Burkholderia mallei* (Bm) are the causative agents of melioidosis and glanders, respectively, and are serious human and animal health hazards in endemic areas throughout the world (Benenson, 1995; Cheng and Currie, 2005; Dance, 2000; Dharakul and Songsivilai, 1999; McGilvray, 1944; Waag and DeShazer, 2004). Both of these pathogenic *Burkholderia* species are listed as category B potential biowarfare agents by the US Centers for Disease Control and Prevention (CDC). Bm has evolved from a clone of Bp (Godoy et al., 2003; Nierman et al., 2004), and the most notable difference between Bm and Bp is that Bm is a highly specialized mammalian pathogen adapted to horses which has not been isolated from the environment, whereas Bp is a soil saprophyte that is an opportunistic pathogen. The genomes of Bm and Bp still share significant nucleotide homology (99%), due to the recent divergence event estimated to have occurred 3.5 million years ago (Lin et al., 2008). However, Bm has undergone extensive genome reduction and reorganizations that mostly have been mediated by insertion sequences (IS) present in the Bm genome in more than 150 copies (Holden et al., 2004; Nierman et al., 2004).

Some species of bacteria use a genetic program, by which certain genes undergo reversible mutations and generate phenotypic diversity in clonal populations. This process is referred to as phase variation and often confers adaptive advantages in various environment (van Belkum et al., 1998; van der Woude and Baumler, 2004). To generate such diversity, many pathogens use Simple Sequence Repeats (SSRs), which consists of tandem arrays of short sequences. These repeats can expand or shrink by the DNA polymerase slippage-based mechanism during DNA replication (Levinson and Gutman, 1987). Variations in number of repeating units in SSRs, located in protein coding regions or upstream regulatory regions, often lead to deactivation or alteration of the associated genes. If these genes are involved in the interactions with the host, such mutations may counteract the host immune response by increasing the antigenic variability of the bacterial population (Levinson and Gutman, 1987; Salaun et al., 2003).

In this study, we investigated the extent to which SSRs are associated with alterations of genes, thereby contributing to the increased diversity in the *Burkholderia* pathogens. We present the comparative genomic analyses across nineteen genomes of Bp and Bm, through which a large number of orthologous genes with SSR-mediated variations were identified and both the genes and their SSRs were characterized.

MATERIALS AND METHODS

Identification of the orthologs among Bp and Bm

In the comparative analyses of the orthologous genes across all Bm and Bp sequenced strains, the 5,799 manually-curated predicted protein sequences from Bp K96243 (Holden et al., 2004) were used as the reference for comparison to all other genomes. These protein sequences were searched against the

Department of Medicine, College of Medicine, Korea University, Seoul 136-705, Korea, [1]Korea Centers for Disease Control and Prevention, Seoul 122-701, Korea, [2]J. Craig Venter Institute, Rockville, MD 20850, USA, [3]The George Washington University School of Medicine, Department of Biochemistry and Molecular Biology, N.W. Washington, DC 20037, USA
*Correspondence: hstanleykim@korea.ac.kr

Received November 18, 2008; accepted November 25, 2008; published online February 20, 2009

Keywords: *Burkholderia mallei, Burkholderia pseudomallei,* pathogen, phase variation, simple sequence repeats

nucleotide sequences of Bm (i.e. ATCC 23344, SAVP1, NCTC 10229, GB8 horse 4, NCTC 10247, 2002721280, FMH, JHU, ATCC 10399) and of Bp (i.e. K96243, 1710a, 1710b, 1106a, 1106b, S13, 1655, 668, 406e, Pasteur 52237) using tblastn (http://blast.wustl.edu). Nucleotide sequences in each genome that were matched to the proteins of Bp K96243 higher than 90% identity over 50% length were identified. For simplicity, these orthologous genes in each genome were referred to by the locus tags in the Bp K96243 annotation regardless of their genomic origin. The average identity of the matched sequences in all genomes was 99%. The genomes used in this study can be obtained from the publicly available Pathema web site (http://pathema.jcvi.org/cgi-bin/Burkholderia/PathemaHomePage.cgi) at the J. Craig Venter Institute (http://www.jcvi.org/) and from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/).

Analysis of the SSR-based variations in the orthologs

We focused on the possibility of phase variation in the protein coding regions of the genes potentially encoding virulence functions. The genes disrupted or deleted in Bm genomes but maintained in Bp were excluded from the analyses, as they are not essential in the virulence of Bm but presumably include mostly those required for environmental survival (Nierman et al., 2004). Those genes that are present in all virulent Bm and Bp strains and depict at least two length variants in 3X-nucleotided SSRs (i.e. those with the repeating units of a multiple number of three, variations in unit number of which do not result in disruptions in the reading frame) were screened, allowing their disruptions or deletions in Bm strains SAVP1, NCTC 10247, and 2002721280 that were tested avirulent in animal models (Detailed strain information is available at : http://pathema.jcvi.org/cgi-bin/Burkholderia/shared/HtmlPage.cgi?page=strains). These sequences were organized based on the variation patterns of the units in SSRs, and each group of orthologous genes were aligned to identify and profile the varied SSR sequences using a multiple sequence alignment program, ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html). To analyze the N-termini of the proteins for the possible presence of the signal peptide, SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) was used.

RESULTS AND DISCUSSION

In this study, we investigated SSR-mediated diversification of the genes in Bp and Bm, by searching for those with at least two unit-number variations in the SSRs. We focused on the genes that are more likely to be essential in Bp and Bm, by confining the analyses to those genes present in all Bp and Bm genomes examined, with the exceptions of three avirulent Bm strains. A total of 210 SSR-varied genes were identified this way. These are a large number found from a single study, and provide a valuable resource to study the SSRs associated with gene diversification and also the type of the genes specifically evolved by this variation system.

Variable SSRs in the coding region of the 210 genes

The SSRs associated with these genes consist of the repeating nucleotide units of a multiple number of three (3X), variations in unit number of which do not result in disruptions in the reading frame. The most prevalent species of these in the 210 genes were found to be the nonamers followed by hexamers and trimers (Fig. 1A). This is intriguing as the nonamers only comprise 13.6% of the 3X-nucleotided SSRs in the Bm ATCC 23344 genome, while the trimers, the most abundant species, comprise 66.7% of them (Nierman et al., 2004). Moreover,

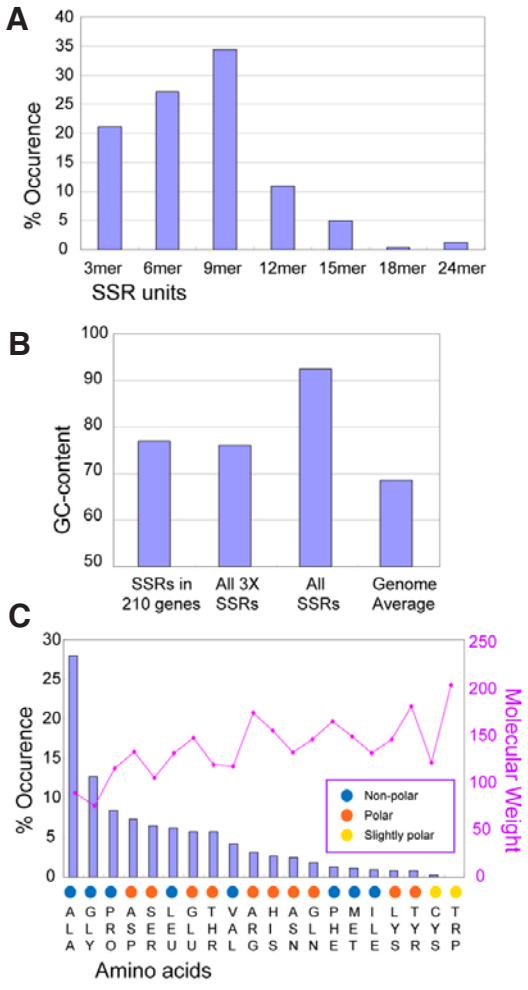

Fig. 1. The varied SSRs associated with the 210 sets of orthologs. (A) Relative abundance of the SSR species. Numbers of occurrence of all seven species of SSRs present at least once in the 210 genes are shown in a bar graph. (B) GC-contents of SSRs. GC-contents of the varied SSRs in the 210 genes are compared in parallel to those from all 3X-nucleotided SSRs, all SSRs, and the genome from Bm ATCC 23344. (C) Relative abundance of the amino acids encoded by the SSRs. Amino acids are denoted by the three letter codes, and the properties of the R-groups of non-polar, polar, and slightly polar are indicated by colored circles. Molecular weights of the amino acids are shown in a graph with connected dots superimposed to the bar graph with their relative abundance within the SSRs.

SSRs of smaller units may allow more sophisticated modifications on proteins than do the nonamers that deal with a rather bulky structural block of three amino acids. The GC content of the varied SSRs associated with the 210 genes was 77.15% on average, which is only slightly higher than the average value calculated from all 3X-nucleotided SSRs (Fig. 1B). These are distinct from the average value of 68.6% in the Bm genome. However they are much closer than is the 92% that is calculated from all other SSRs present in the genome. This GC-content value of the SSRs in the 210 genes reflects the unbalanced codon preference of the amino acids with relatively smaller and nonpolar R groups are preferentially represented in these repeats, with Alanine, Glycine, and Proline being the

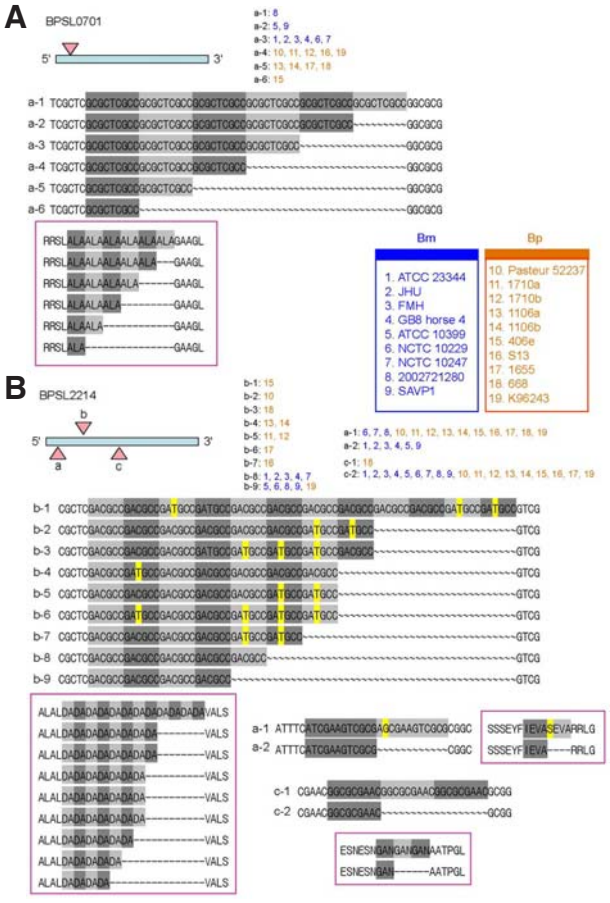

Fig. 2. SSR-mediated diversification of two orthologous genes across nineteen genomes of Bp and Bm. The locus names for the orthologous genes are adapted from those of Bp K96243 (Holden et al., 2004). (A) BPSL0701 with repeat-unit variations. The approximate position of the varied SSR is indicated with an arrowhead on the box representing the gene. All the SSR patterns found is shown from the longest at the top. Translated sequences for the corresponding regions are shown in a box below the nucleotide sequences. Strains of Bp and Bm are denoted by the numbers that correspond to the strains listed in two boxes labeled Bm and Bp. (B) BPSL2214 with three varied-SSRs. Three positions of the SSRs on the gene are indicated by the arrow heads labeled a, b, and c. The SSR patterns of these and their translated sequences are displayed close to each other. Nucleotides with point mutations in the SSRs are highlighted yellow.

most prevalent (Fig. 1C). Consistent with this observation, the largest and slightly polar amino acid Tryptophan was the least frequently encoded by these repeats. Cysteine, that can severely alter the conformation of proteins by forming sulfur bridges, was the second least encoded amino acid by these SSRs. This suggests that there has been a selection process for this group of SSRs to generate alterations in genes which do not significantly cause deleterious effects in the resultant proteins.

We found that there are basic patterns in the SSR-based variations in the 210 genes. The simplest type of variations can be exemplified by BPSL0701 that codes for a putative exported protein (Fig. 2A) (orthologs are referred to by the annotation of Bp K96243). This gene contains a single SSR with the unit of 5′-GCGCTCGCC-3′ at the region close to the 5′-end of the gene. Deletions and insertions of the repeat units alter the protein by the unit of Ala-Leu-Ala. On the other hand, there also are the genes that have multiple variable SSRs. Some contain unrelated SSRs, while others have those duplicated and diverged from an SSR. Although derived from an SSR, these appeared to go through unit number variations independent of each other. Multiple unrelated SSRs are present in BPSL2214, which has as many as three different SSRs with the units of, 5′-A(T/g)CGAAGTCGCG-3′, 5′-GA(C/t)GCC-3′, and 5′-GGCGC-GAAC-3′ (the lower case letters indicate the nucleotides in the position that are less frequent) (Fig. 2B). In contrast, BPSS1307, coding for a putative ATP-binding ABC transporter protein, represents the ones containing SSRs duplicated and diverged from an SSR (Supplemental Fig. S1). This gene contains SSRs of 5′-TCGACCG(C/t)T-3′ and 5′-TC(G/a)GCCG(C/a)T-3′ that code for ST(A/v) and SA(A/d), respectively. Some SSRs contain point mutations in the units that are either synonymous or non-synonymous. Genes BPSL2214 and BPSS1307 show such mutations in their SSRs. In BPSL2214, both synonymous [(GAC (Aspartic acid) → GAT (Aspartic acid)] and non-synonymous [ATC (Isoleucine) → AGC (Serine)] changes are present. These mutations may be part of the fine-tuning of the adaptive process of the resultant proteins to the environment.

Genes of putative phase variation

Among the most noticeable characteristics of the 210 genes is that many of them (62%) contain predicted leader sequences for secretion at the 5′-end, and based on the annotation for Bp K96243 (Holden et al., 2004), a large portion (ca. 81%) of the group appears to encode a protein localized to the surface or secreted (Fig 3; Supplemental Table S1). These suggest that they may directly interact with the host factors during pathogenesis or variable environmental factors. This justifies the need to vary their antigenic properties to evade the host immune systems. Besides these genes, there also are other variable protein products which function inside the cell. While they may not directly interact with the host factors, their diversity may also confer increased adaptability to the bacteria in the ever-changing environment.

Based on the SSR variation patterns, these genes fall into three types: 1) those varied in both Bp and Bm, 2) those varied in one species, and 3) the genes not varied within each species but exhibiting distinct patterns between Bp and Bm. We found a total of nineteen genes for the first category (Fig. 3; Supplemental Table S1). Among the most varied genes in this group is BPSL2214, which has as many as nine length variants caused by three different SSRs (Figs. 2B and 3). The second group of genes comprises the largest group of 164 genes. Sixty two of these have at least three length variations in the SSR (for all data, see Supplemental Table S1). We expect that these genes may be associated with more species (or niche)-specific activities. Bp strains are of two geographical origins; strains 1655 and 668 were isolated from Australia, while the others are all from the Southeast Asia. We found a large number of genes distinctive between the two groups (Fig. 3; Supplemental Table S1). Most notable group of genes are the ones that are varied in strains 1655 and/or 668, while all others including Bm stains have the same SSR patterns. A single gene BPSS1864, which codes for a putative AraC-family regulatory protein, showed the same variation pattern in both strains 1655 and 668, an insertion of an SSR unit of 5′-GCGGCAGCGCCGACG-3′ (Supplemental Table S2). Others were with distinct SSR variations only in one of these two Australian strains, and there were more of those in 668 than were in 1655. Interestingly, there were signifi-

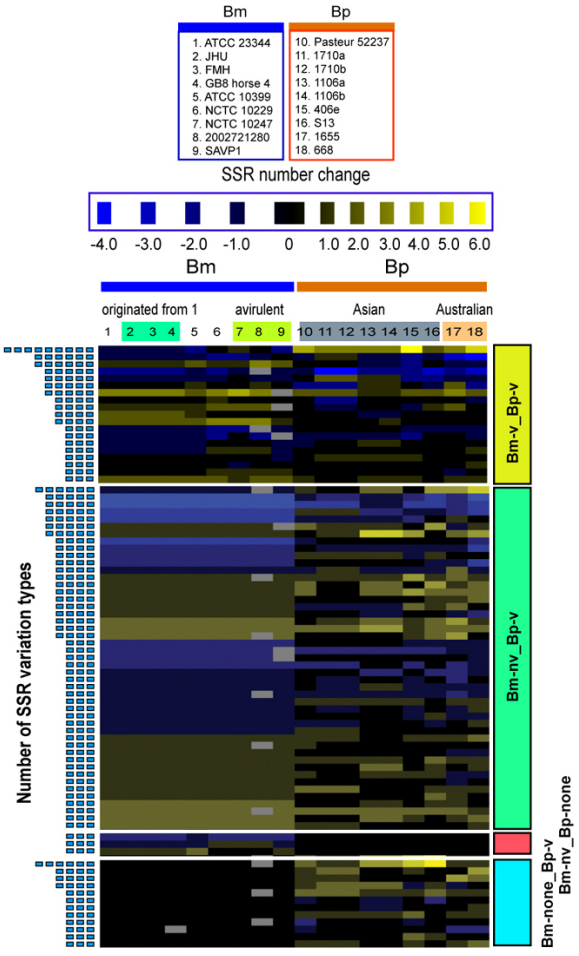

Fig. 3. Genes with varied SSRs in Bp and Bm. The genes present in Bp K96243 are used as standards, and insertions or deletions of the repeat units in the orthologs in all other Bp and Bm strains are relative to them. The spectrum of two colors of the solid boxes at the top indicate insertions (+) or deletions (-) of SSR units, while distinct level of the intensity represents each of the SSR length variant, starting from ± 1 for the first variant that is the closest to the standard. The total number of SSR variation types in each gene is shown with corresponding number of blue boxes on the left of the heat map figure. Gene annotation for Bp K95243 was used for all the orthologs in the genomes of Bp and Bm. Bm-v_Bp-v, the genes varied in both Bm and Bp; Bm-nv_Bp-v, the genes that did not vary in Bm but distinct from the standards in Bp K96243, but did in Bp; Bm-v_Bp-none, the genes that varied in Bm but did not in Bp; Bm-none_Bp-v, genes not varied in Bm but with the same variation patterns as the standards from Bp K96243 but varied in Bp. All the original data in the figure as well as the others not displayed can be found in the Supplemental Table S1.

cantly more genes varied in Bp than are varied in Bm, and this may reflect wider range of environments that Bp encounters, in contrast to a single host, the equines, with which Bm is exclusively associated. This may also be partly due to the recent divergence of Bm (~3.5 million years) (Lin et al., 2008) from a clone of Bp (Godoy et al., 2003; Nierman et al., 2004). Lastly, there are the genes not varied within each species but show distinct patterns between Bp and Bm. These may not require frequent changes as much as the others, and perhaps simply reflect clonal traits. Twenty-seven genes in this group are listed in the Supplemental Table S1.

On the other hand, we did not find recent variations in the SSRs present in three Bm strains, JHU, FMH, and gb8-horse that were originated from the strain ATCC 23344 (Romero et al., 2006); the first two were obtained after a passage in a human, while the last strain was recovered from an infected horse. Nor did we find them in three avirulent Bm strains NCTC 10247, 2002721280, and SAVP1, which may be under relaxed evolutionary pressure to keep the virulence genes functional. In contrast, a single gene BPSS0524 was found varied in an SSR with 5′-CAG-3′ (Supplemental Table S2) between Bp 1710a and Bp 1710b that were isolated from the same patient but with a 3 years of interval between the initial infection and the relapse (Romero et al., 2006). Intriguingly, this same gene, coding for a conservative hypothetical protein, also differs between Bp 1106a and 1106b that were obtained from another patient with sampling points of three years apart. Taken together, results from these three groups of strains suggest that SSR variations in genes are rare at least under the environmental changes applied.

CONCLUSION

Our analysis provides the detailed view of the SSR system that is used to diversify coding regions of genes in *Burkholderia* pathogens Bp and Bm. Another mechanism of increasing diversity is through phase variation of the promoters, controlling the on and off or modulation of the expression of certain genes. Different kinds of SSRs instead of 3X-nucleotided ones we studied are expected to be used in this promoter diversification process, and studies in this area will be of importance in *Burkholderia*. Equally important is the detailed studies with the 210 genes and their proteins discovered from this study. As many of these proteins may be involved in close interactions with the host, they are of significance for basic research as well as for the development of diagnostics, therapeutics, and vaccines against these important *Burkholderia* pathogens.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This work was supported by the grants M102KK010005-08K1101-00530 from the 21C Frontier Microbial Genomics and Applications Center Program and RO1-2008-000-11047-0 from the Basic Research Program, both of which from the Ministry of Education, Science and Technology in the Republic of Korea, and by the Korea Foundation for International Cooperation of Science and Technology through a grant provided by the Korean Ministry of Science and Technology in K20601000002-07E0100-00240 to H.S.K. Sequencing and annotation of most of the Bp and Bm strains was supported by NIAID N01-AI30071 to W.C.N.

REFERENCES

Benenson, A. (1995). Control of communicable diseases manual (Washington, DC, American Public Health Association).
Cheng, A.C., and Currie, B.J. (2005). Melioidosis: epidemiology, pathophysiology, and management. Clin. Microbiol. Rev. *18*, 383-416.
Dance, D. (2000). Ecology of *Burkholderia pseudomallei* and the interactions between environmental *Burkholderia* spp. and human-animal hosts. Acta Trop. *74*, 159-168.
Dharakul, T., and Songsivilai, S. (1999). The many facets of melioidosis. Trends Microbiol. *7*, 138-140.

Godoy, D., Randle, G., Simpson, A., Aanensen, D., Pitt, T., Kinoshita, R., and Spratt, B. (2003). Multilocus sequence typing and evolutionary relationships among the causative agents of Melioidosis and Glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. J. Clin. Microbiol. *41*, 2068-2079.

Holden, M.T.G., Titball, R.W., Peacock, S.J., Cerdeño-Tárraga, A.M., Atkins, T., Crossman, L.C., Pitt, T., Churcher, C., Mungall, K., Bentley, S.D., et al. (2004). Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. Proc. Natl. Acad. Sci. USA *101*, 14240-14245.

Levinson, G., and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. *4*, 203-221.

Lin, C.H., Bourque, G., and Tan, P. (2008). A comparative synteny map of *Burkholderia* species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. Mol. Biol. Evol. *25*, 549-558.

McGilvray, C. (1944). The transmission of glanders from horse to man. Can. J. Public Health *35*, 268-275.

Nierman, W.C., DeShazer, D., Kim, H.S., Tettelin, H., Nelson, K.E., Feldblyum, T., Ulrich, R.L., Ronning, C.M., Brinkac, L.M., and Daugherty, S.C. (2004). Structural flexibility in the *Burkholderia* genome. Proc. Natl. Acad. Sci. USA *101*, 14246-14251.

Romero, C., DeShazer, D., Feldblyum, T., Ravel, J., Woods, D., Kim, H.S., Yu, Y., Ronning, C., and Nierman, W. (2006). Genome sequence alterations detected upon passage of *Burkholderia mallei* ATCC 23344 in culture and in mammalian hosts. BMC Genomics *7*, 228.

Salaun, L., Snyder, L.A.S., and Saunders, N.J. (2003). Adaptation by phase variation in pathogenic bacteria. Adv. Appl. Microbiol. *52*, 263-301.

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. MMBR *62*, 275 - 293.

van der Woude, M.W., and Baumler, A.J. (2004). Phase and antigenic variation in bacteria. Clin. Microbiol. Rev. *17*, 581-611.

Waag, D.M., and DeShazer, D. (2004). Glanders: New Insights into an Old Disease (New Jersey: Humana Press).