

# **Genome Analysis of *Burkholderia pseudomallei***

Xie Chao

Submitted to the Department of Biochemistry, National University of Singapore  
in partial fulfillment for the degree of Bachelor of Science (Honours)

2003/2004

## **ACKNOWLEDGEMENT**

I would like to express my sincere appreciation to my supervisors, Dr Chua Kim Lee and Dr Tan Tin Wee, for their guidance, patience, and frank advises throughout the entire project

## ABSTRACT

The *Burkholderia pseudomallei* genome sequencing was completed by the Sanger Institute on May 2003. It comprises of two replicons, known as chromosomes 1 and 2, respectively. The genome annotation is yet to be completed and genome scale analysis has not been done. The aim of this study was to annotate and analyze the *B. pseudomallei* genome. The genome was annotated by homology and conserved domain search and the genes thus identified were then classified and stored into a database called *Burkholderia pseudomallei* Genome Database (<http://origin-sgi.bic.nus.edu.sg/~xiechao/query.html>). By applying the annotation database to the analysis of the *B. pseudomallei* genome, many characteristics of the genome were revealed. The analysis showed that the larger replicon carries a complete set of genes encoding all the essential pathways, while the smaller replicon encode only very few complete pathways. This suggests that the smaller replicon may be extrachromosomal and probably dispensable. 49.4% of the genes in the larger replicon share high homology with genes of *Ralstonia solanacearum*, while the phylogenetic origins of the smaller replicon is much more diverse. Additionally, the origin of replication on the smaller replicon resembles that of a plasmid. Thus, the smaller replicon could be a megaplasmid which might have been acquired through gene transfer mechanisms. The usefulness of the *B. pseudomallei* database was demonstrated by its application to the discovery of seven putative pathogenicity islands in *B. pseudomallei*.

# Table of Contents

<b>Abstract</b>	<b>3</b>
<b>Table of Contents</b>	<b>4</b>
<b>List of Tables</b>	<b>6</b>
<b>List of Figures</b>	<b>7</b>
<b>List of Programs or Scripts</b>	<b>8</b>
<b>Abbreviations and symbols Used</b>	<b>9</b>
<b>Chapter 1 Introduction</b>	<b>10</b>
<b>Chapter 2 Creation of the <i>Burkholderia pseudomallei</i> Genome Database</b>	<b>12</b>
2.1 Introduction	12
2.2 Materials and Methods	13
2.3 Results	17
2.3.1 Table design for the <i>Burkholderia pseudomallei</i> genome database	17
2.3.2 Data for the <i>Burkholderia pseudomallei</i> genome database	19
2.3.3 Query and manipulation scripts for the database	21
2.3.4 An example of application of the database, RND efflux systems identification	24
2.4 Discussions	25
<b>Chapter 3 Analyses of the <i>Burkholderia pseudomallei</i> Genome</b>	<b>27</b>
3.1 Introduction	27
3.2 Materials and Methods	27
3.3 Results	30
3.3.1 Paralogous groups in <i>B. pseudomallei</i>	30
3.3.2 Comparative genomics	31
3.3.3 Origin of replication in <i>B. pseudomallei</i>	34
3.3.4 Functional categories in <i>B. pseudomallei</i>	36
3.3.5 Pathway identifications in <i>B. pseudomallei</i>	38
3.3.6 Transcriptional regulators in <i>B. pseudomallei</i>	42
3.4 Discussions	44

<b>Chapter 4</b>	<b>Detection of Pathogenicity Islands and Putative Alien Gene Clusters in <i>Burkholderia pseudomallei</i></b>	<b>47</b>
4.1	Introduction	47
4.2	Materials and Methods	49
4.3	Results	51
4.3.1	Implementation of Karlin's 5 criteria	51
4.3.2	Prediction of pathogenicity islands in <i>B. pseudomallei</i>	53
4.3.3	Prediction of putative alien gene clusters in <i>B. pseudomallei</i>	58
4.4	Discussions	60
<b>Chapter 5</b>	<b>Discussions and Conclusion</b>	<b>62</b>
<b>References</b>		<b>66</b>

## List of Tables

Table 2.1	Summary of records in table “gene_cog”, “gene_pfam”, and “gene_smart”
Table 2.2	Known and putative RND multidrug efflux pumps in <i>B. pseudomallei</i>
Table 3.1	General Features of the <i>B. pseudomallei</i> genome
Table 3.2	Paralogous groups in <i>B. pseudomallei</i>
Table 3.3	The top 10 organisms with best similarity with <i>B. pseudomallei</i> genome
Table 3.4	Comparison of <i>B. pseudomallei</i> ORFs on the two chromosomes with other organisms
Table 3.5	The replication proteins on chromosome 2.
Table 3.6	Functional categories of predicted ORFs
Table 3.7	Amino Acid biosynthesis pathways in <i>B. pseudomallei</i> .
Table 3.8	Nucleotide biosynthesis/salvage pathways in <i>B. pseudomallei</i> .
Table 3.9	Energy production and conversion pathways in <i>B. pseudomallei</i> .
Table 3.10	Electron transport and oxidative phosphorylation in <i>B. pseudomallei</i> .
Table 3.11	Coenzyme biosynthesis pathways in <i>B. pseudomallei</i> .
Table 3.12	Replication/Transcription/Translation basal machinery in <i>B. pseudomallei</i> .
Table 3.13	Other pathway or machinery in <i>B. pseudomallei</i> .
Table 3.14	Transcriptional regulator motifs in <i>B. pseudomallei</i> .
Table 4.1	Files in the PAI prediction program package
Table 4.2	The PAI candidates on chromosome 1
Table 4.3	The PAI candidates on chromosome 2
Table 4.4	Putative alien (pA) genes in <i>B. pseudomallei</i> genome
Table 4.5	Putative alien (pA) gene clusters in chromosome 1
Table 4.6	Putative alien (pA) gene clusters in chromosome 2

## List of Figures

- Figure 2.1 The 16 tables in the *Burkholderia pseudomallei* Genome Database.
- Figure 2.2 A simplified ER diagram showing the overall structure of the *Burkholderia pseudomallei* Genome Database
- Figure 2.3 Sample output of a genome database query (query by name keyword “dnaA”)
- Figure 3.1 Comparison of the ORFs of the two chromosomes with other organisms
- Figure 3.2 The GC skew analysis of Chromosome 1 and 2.
- Figure 3.3 The origin of replication and the replication proteins of chromosome 2.
- Figure 4.1 User interface of PAI prediction package for PAI prediction.
- Figure 4.2 User interface of PAI prediction package for the pA gene clusters prediction.
- Figure 4.3 PAI prediction on chromosome 1
- Figure 4.4 PAI prediction on chromosome 2

## List of Programs or Scripts

### PAI prediction

pai.java, gcodon.java, pA.java, readFASTA.java, ORFreader.java

### BLAST, comparative genomics, and paralogous group identifications

NCBI BLAST: Blaster.java, doBlast.java, errorBlast.java, ORFreader.java

Local BLAST: Blaster.pl, format.pl, splitFasta.pl

Analyze Result: readBlast.java, pickName.java, extractCompare.java, extractCompare2.java,  
analyzeParalog.java, countGroup.java

### Conserved Domain Database Search

cdd.java, errorCDD.java, readFASTA.java, extractResult.java

### Origin of replication

origin.java, readFASTA.java, drepeat.java

### CGI scripts for the web interface of the genome database

acc.pl, query\_by\_class.pl, query\_go\_term.pl, query\_by\_cog.pl, show\_detail.pl, blast.pl,  
query\_by\_go.pl, ccd.pl, query\_by\_keywords.pl, togb.pl, do\_update.pl, query\_by\_pfam.pl, update.pl,  
open\_go.pl, query\_by\_position.pl, query.pl, query\_by\_sql.pl

### Database maintain scripts

toFasta.pl, updateSequence.pl

### Pathway prediction

pathwayCOG.java, pathwayGene.pl

## Abbreviations and symbols Used

CDD	Conserved Domain Database
CGI	Common Gateway Interface
COG	Clusters of Orthologous Groups of proteins
GO	Gene Ontology
GUI	Graphical User Interface
Kbp	Kilo base pairs
Mbp	Mega base pairs
ORF	Open Reading Frame
pA	Putative Alien
PAI	Pathogenicity Island
Pfam	Protein Family Database

## CHAPTER 1. INTRODUCTION

*Pseudomonas pseudomallei* was first described in 1912 by Whitmore and Krishnaswami (Whitmore and Krishnaswami 1912). It was transferred into a new genus, *Burkholderia* which was proposed in 1992 by Yabuuchi *et al* (Yabuuchi *et al.* 1992).

*B. pseudomallei* is a free-living gram-negative bacillus which was found most frequently in soil, stagnant water, and rice paddies in areas where it is endemic (Sexton *et al.* 1993). It is the causative agent of melioidosis, an important infectious disease in S.E. Asia, particularly N.E. Thailand and northern Australia (Yabuuchi and Arakawa 1993). Most of the recent research has been focused on the identification of virulence factors of *B. pseudomallei* and the development of methods for early and definitive diagnosis and treatment of melioidosis. These virulence factors are often encoded within clusters of multiple genes that are involved in the process of pathogenesis. Such gene clusters are called pathogenicity islands (PAIs) (Hacker and Kaper 2000).

*B. pseudomallei* is a beta-proteobacterium, a group of bacteria whose genomic organization is still poorly characterized (Salanoubat *et al.* 2002). It has a very large genome, which is larger than most other bacteria. The genome comprising two replicons of 4Mbp and 3Mbp was completed on May 2003 by the Wellcome Trust Sanger Institute, UK. However, annotation of the sequences is still ongoing and the full genome annotation is not published yet.

In this project, the genome was systematically annotated, analyzed, and stored in an Internet accessible relational database called the *Burkholderia pseudomallei* Genome Database. In addition, various query and manipulation scripts were also developed to facilitate ease of interrogation and data mining of the *B. pseudomallei* Genome Database. This database was further used for the analysis of the *B. pseudomallei* genome.

Several aspects of the genome were analyzed, including its probable evolution history, origin of replication, functional categories, pathways, transcriptional regulators, and putative PAIs. These results would be helpful for the understanding of the organization and evolution of the *B. pseudomallei* genome. They were also used to differentiate the two replicons. The analysis suggests that the smaller replicon is probably extrachromosomal as it has many features of megaplasmid.

A program for identifying putative PAIs was developed. This package used the criteria that were proposed by Karlin to predict putative PAIs (Karlin 2001). Using this program, 7 putative PAIs were identified in the *B. pseudomallei* genome. These data might be useful for the discovery of genes involved in *B. pseudomallei* virulence and pathogenesis.

## **CHAPTER 2. CREATION OF THE *BURKHOLDERIA PSEUDOMALLEI* GENOME DATABASE**

### **2.1 Introduction**

As of March 2004, of the 335 microbial genomes that are being sequenced, only 155 microbial genomes have been completely sequenced, annotated and deposited in Genbank (NCBI). Of those not completed, genome annotation is a key bottleneck, without which the raw genome sequence data does not provide biologically meaningful information.

Annotation of the genome sequences is the bridge between the sequence data and the biology of the organism (Stein 2001). The aim of the annotation is to identify the features of the genome, particularly the genes, their regulator elements and their products.

The sequence of the entire *Burkholderia pseudomallei* genome comprising two circular replicons of 4Mbp and 3Mbp was completed on May 2003 by the Sanger Institute. However, annotation of the sequences is still ongoing and the full genome annotation is not published yet. In this chapter, we describe a process for rapidly annotating this microbial genome, and the database used to organize the information.

The genome sequence data of *B. pseudomallei* used in this project were produced by the Sequencing Group of the Sanger Institute. The data were systematically annotated, analyzed, and stored in an Internet accessible database called the *Burkholderia pseudomallei* Genome Database. Various query and manipulation scripts were also developed to interact with the *B. pseudomallei* Genome Database.

The database was constructed in the following way. Firstly, the open reading frames (ORFs) were predicted from the genome sequence using standard techniques. Each of the predicted genes was then analyzed, first by homology search, followed by protein conserved

domain search. The protein-level annotation was mostly dependent on the homology search, while the functional class- or pathway-level annotation was dependent on the protein family domain analysis which used both Protein Family Database (Pfam) (Bateman *et al.* 2002) and Clusters of Orthologous Groups of proteins (COGs) (Tatusov *et al.* 2001). The annotated genes were classified by the COG's functional categories (Tatusov *et al.* 1997) and the Gene Ontology (GO) vocabulary (Ashburner *et al.* 2000).

## **2.2 Material and Methods**

### **2.2.1 Materials**

*Burkholderia pseudomallei* genome sequences of strain K96243 was used to construct the database in this project. The sequence data were produced by the *Burkholderia pseudomallei* Sequencing Group at the Sanger Institute and can be obtained from <ftp://ftp.sanger.ac.uk/pub/pathogens/bps/>. The sequence was downloaded on 2 July 2003. No subsequent updates were available.

Both Java and Perl were used for programming. The version of Java that was used in this project is J2SE 1.4.2 (<http://java.sun.com>), and the version of Perl is ActivePerl 5.8.0.806 (<http://www.activestate.com/>). Additional modules (DBI and DBD::mysql) for Perl were installed in order to interact with the database through Perl scripts. All the source codes of the programs written in this project are available at <http://origin-sgi.bic.nus.edu.sg/~xiechao/code/>.

The relational database management system chosen for this project was the freely accessible MySQL (<http://www.mysql.org/>). Internet access to this database is via a user-friendly web interface using an Apache webserver (<http://www.apache.org/>) and Perl-based Common Gateway Interface (CGI) scripts.

### **2.2.2 Overview of the Database Construction Strategy**

The open reading frames were firstly predicted for the protein coding genes (details are in

section 2.2.3). Then the translated products of the ORFs were analyzed by BLAST (Sections 2.2.4, 2.2.5) and Conserved Domain Database (CDD) search (Section 2.2.6). The ORFs were then assigned one or more functional classes according the domain search results (Section 2.2.7). They were also mapped to Gene Ontology terms (Section 2.2.8). The non-coding RNA species (tRNA and rRNA) were analyzed too (Section 2.2.9).

The relational database tables were designed and created in a semi-normalised form (Section 2.2.10), and the analysis results were loaded. After construction of the database, various scripts were written for query or maintenance of the database (Section 2.2.11).

### **2.2.3 Open Reading Frames Prediction for the Protein Coding Genes**

The ORFs were firstly predicted by GeneMark (Besemer *et al.* 2001). GeneMark web server is available at <http://opal.biology.gatech.edu/GeneMark/>. Several low ORF density regions (no ORF for 10 Kbp region) from the GeneMark result were analyzed manually.

### **2.2.4 BLAST against NCBI and Local BLAST Server**

The translated protein products were first analyzed by BLAST on NCBI BLAST server to find the protein homologs (Altschul *et al.* 1997). The BLAST program used was **blastp**, and the database used was “nr”. The expect value was 1e-5. The first BLAST search date for the database construction was 19 July 2003. The latest update of the homologs is by a new **blastp** search against all the proteomes of all the 155 fully sequenced microbial organisms on 25 Feb 2004 on a local machine.

A Java program was written in order to carry out the large scale BLAST. There are 4 files in the program. **Blaster.java**, **doBlast.java** and **ORFReader.java** were used to read genome sequence and ORF information file, send each BLAST request to NCBI BLAST server, and get back the results. Any error during the BLAST search, either by network or the server, was recorded in an error log file. The **errorBlast.java** reads the error log and repeats the failed

BLAST. The speed of the BLAST search was limited in order to avoid overwheel of the NCBI BLAST server.

For local BLAST, Perl script **blaster.pl** was written. The BLAST engine was the stand-alone BLAST programs downloaded from NCBI.

### 2.2.5 Extraction of Information from BLAST Results

Two Java programs were written for processing the BLAST results. **readBlast.java** reads all the BLAST results html files from a directory, and extracts The homolog names if they can satisfy two criteria.

1. The expect value must be less than  $1e-5$  (valid in the BLAST settings);
2. The difference of lengths between the homolog and query protein must be less than half of the query length.

The homologs were recorded in a file, and the file was processed by another Java program, **pickName.java**, in order to pick the homologs name from the output of **readBlast.java**.

### 2.2.6 Conserved Domain Database Search (Pfam, Smart, and COG)

The ORFs were analyzed for protein conserved domains at NCBI Conserved Domain Database (CDD) (Marchler-Bauer *et al.* 2003). A Java program was written to carry out the search. There were three files in the program: **cdd.java**, **readFASTA.java** and **errorCDD.java**. The **readFASTA.java** reads the ORFs sequences from a FASTA file. The **cdd.java** sends request to NCBI CDD server (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). The database was CDD version 1.65. This database contains Pfam version 11.0 (Bateman *et al.* 2002), Smart version 4.0 (Letunic *et al.* 2004), and COG version 1.0 (Tatusov *et al.* 2001). The expect value was set to  $1e-4$ . The **errorCDD.java** repeats the search according the the error log file. Another Java program, **extractResult.java**, was used to extract the domain or family accession number (Pfam, Smart, and COG) from the directory of CDD search results.

### **2.2.7 Functional Categories Assignment**

The structure of protein functional categories follows the COG database format (Tatusov *et al.* 1997). The correspondence of the COG domain to functional classes was obtained from NCBI ftp server (<ftp://ftp.ncbi.nih.gov/pub/COG/>). Then all genes that have COG domains were assigned to one or more functional classes through the link between COG domain and functional categories.

### **2.2.8 Gene Ontology Assignment**

Gene Ontology (Ashburner *et al.* 2000) was used to classify the ORFs with a more detailed manner than the simple functional categories. The vocabulary of GO used was the October 2003 release. The assignment was carried out using the CDD Search results (Pfam and Smart) and the maps from Pfam and Smart to GO term. The maps were constructed by Nicola Mulder, and downloaded from <http://www.geneontology.org/external2go/>. A new term, virulent factors, were added into the GO vocabulary. The child terms of the virulent genes were collected according the book *Bacterial Protein Toxins* (Burns *et al.* 2003).

### **2.2.9 tRNA and rRNA analysis**

tRNAscan-SE 1.21 (Lowe and Eddy 1997) was used to predict the tRNAs in the genome. The program was downloaded from <ftp://ftp.genetics.wustl.edu/pub/eddy/software/>. The analysis was carried out on a local machine. The rRNA sequence was predicted by **blastn** using known rRNA as query.

### **2.2.10 Database Design and Construct**

The *B. pseudomallei* Genome Database was designed and constructed on a MySQL server (<http://www.mysql.org>). The version of MySQL is 4.0.15. Tables for the genome sequences and analysis results were designed and normalized as appreciate. These information in the database contains the genes' location, names, and BLAST homologs, Pfam/Smart/COG

domain matches, functional categories, and Gene Ontology analysis results.

### **2.2.11 Database Query and Update Scripts Development**

CGI scripts were written in Perl for the query and update of the *B. pseudomallei* Genome Database. Several types of gene queries were implemented, such as query by gene accession, by name keywords, by location, by COG, by Pfam, by functional class, and so on.

A BLAST web server was also set up for the *B. pseudomallei* Genome Database in order to do sequence search. The **www-BLAST** programs were downloaded from NCBI ftp server (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/>). Both the DNA and protein sequences of the *B. pseudomallei* Genome Database were formatted for the BLAST web server.

For easier maintain and update of the database, Perl scripts were also written for the update of the database through web interface.

### **2.2.12 Potential RND efflux systems prediction**

Potential RND efflux pumps were discovered from the *B. pseudomallei* Genome Database using COG, Pfam domains, and GO terms. As a complement, known RND efflux systems were used to do **blastp** against the *B. pseudomallei* Genome Database. All the possible candidate genes were examined manually.

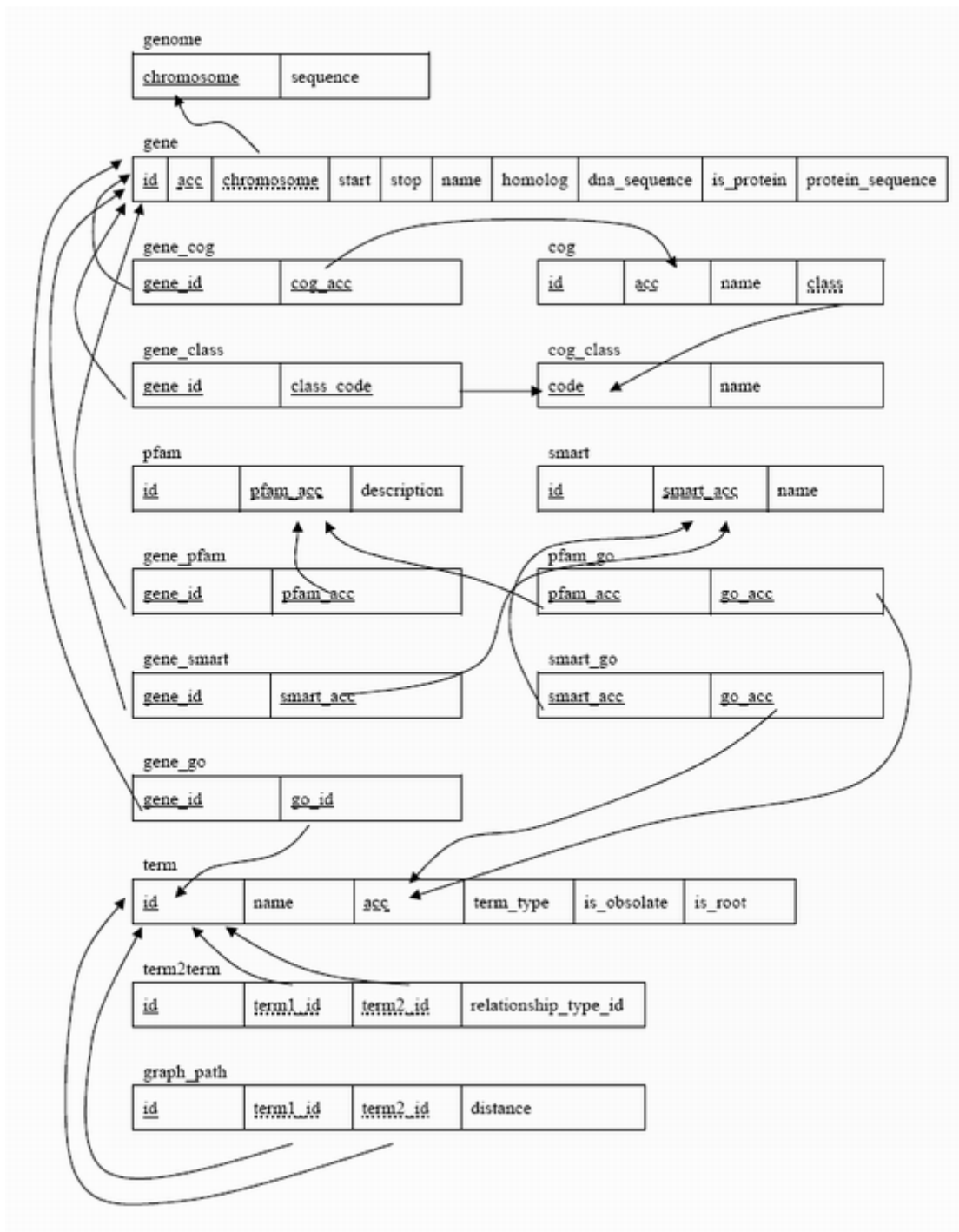
## **2.3 Results**

### **2.3.1 The Table Design of the *Burkholderia pseudomallei* Genome Database**

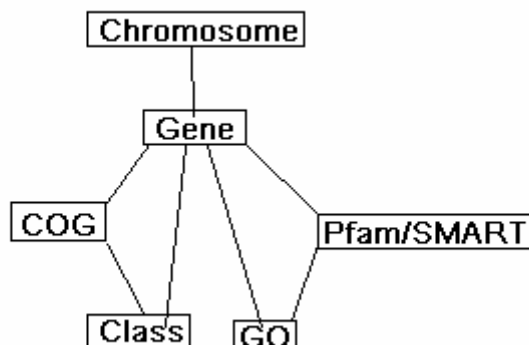
16 tables were constructed in the *B. pseudomallei* Genome Database (Figure 2.1). The overall structure of these tables can be shown in a simplified ER diagram (Figure 2.2). The entity “Gene” (table “gene”) is the central of the database. It contains the basic information for the genes, such as the location, accession number, name, and homolog of the genes. It also points to the “Chromosome” (table “chromosome”), which can be used for both coding and non-coding region analysis. More information, such as domain or motif, of the genes can be

found in the “COG” (table “gene\_cog”, “cog”) and “Pfam/Smart” (table “gene\_pfam”, “pfam”, “gene\_smart”, and “smart”). The genes are also classified into functional categories, such as “Class” (table “cog\_class”, “gene\_class”), and more detailed ontology “GO” (table “pfam\_go”, “smart\_go”, “gene\_go”, “term”, “term2term”, and “graph\_path”).

**Figure 2.1 The 16 tables in the *Burkholderia pseudomallei* Genome Database.**  
 The notations: Primary key; Unique key; Foreign key (reference to →).



**Figure 2.2 A simplified ER diagram showing the overall structure of the *Burkholderia pseudomallei* Genome Database**



### **2.3.2 Data for the *B. pseudomallei* Genome Database**

The data for every column of each table are described in following sections.

#### **2.3.2.1 Table “chromosome”**

The two chromosomes’ sequences were loaded into the table “chromosome”.

#### **2.3.2.2 Table “gene”**

5679 ORFs for the protein coding genes, 12 ribosomal RNAs, and 59 transfer RNAs were predicted (column “chromosome”, “start”, ”stop”, “is\_protein”). 5046 of the ORFs have homologs (column “homolog”). The putative names of these genes were picked up from their function-known homologs (column “name”). The column “dna\_sequence” and “protein\_sequence” was loaded with the genes’ DNA sequence and protein sequence (for protein coding genes).

#### **2.3.2.3 Table “cog”, “cog\_class”**

The data for the table “cog” and “cog\_class” were obtained from NCBI’s ftp server. The table “cog” contains 4873 COG domains, with their accession number (column “acc”), name (column “name”), and class code (column “class\_code”). The table “cog\_class” contains the 25 functional class codes (column “code”) and explanation of the code (column “name”).

#### 2.3.2.4 Table “*pfam*”, “*smart*”, “*pfam\_go*”, “*smart\_go*”

The Pfam and SMART domain or motif information were obtained from their ftp servers. There are 6190 Pfam and 665 SMART records in the database, with their accession numbers (column “acc”) and names (column “name”, “description”).

The table “*pfam\_go*” and “*smart\_go*” contains the information the map from Pfam/SMART domain (column “*pfam\_acc*”, “*smart\_acc*”) to the GO term (column “*go\_acc*”). Those data were collected from GeneOntology.org.

#### 2.3.2.5 Table “*gene\_cog*”, “*gene\_pfam*”, “*gene\_smart*”

The data for the three tables were from the CDD search results. Each of them has two columns, the gene ID (column “*gene\_id*”) and COG/Pfam/SMART accession number (column “*cog\_acc*”, “*pfam\_acc*”, and “*smart\_acc*”). The records of the three table are summarized in Table 2.1.

**Table 2.1 Summary of records in table “*gene\_cog*”, “*gene\_pfam*”, and “*gene\_smart*”**

Table	Total records	Distinct genes	Distinct Domains
<i>gene_cog</i>	10959	4332	2403
<i>gene_pfam</i>	5493	4161	1727
<i>gene_smart</i>	596	491	110

#### 2.3.2.6 Table “*gene\_class*”

The table “*gene\_class*” contains the information of the genes’ functional categories. This table (column “*gene\_id*”, “*class\_code*”) is a joint table of table “*gene\_cog*” and “*cog*” by the equality of column “*cog\_acc*”. 4332 genes were categorized into 20 functional classes.

Although all the information in the table “*gene\_class*” can be obtained from table “*gene\_cog*” and “*cog*”, this table is useful in the means of query efficiency. With this table, it

is no need to do the join operation everytime when the genes' functional categories are queried.

#### 2.3.2.7 Table “term”, “term2term”, “graph\_path”

The three tables (table structure and data) are part of the GO official release. The table “term” describes the basic information for the GO terms, such as the term identifier (column “id”), accession number (column “acc”), name (column “name”), and others (column “is\_root”). The table “term2term” describes the “parent-child” relationship between GO terms, and thus defines the GO tree structure. The table “graph\_path” is a calculated product from table “term2term”, and it describes the “is a” relationship for GO terms. This table is used to improve query efficiency, as table “gene\_class”.

#### 2.3.2.8 Table “gene\_go”

The table “gene\_go” have two columns, “gene\_id” and “go\_id”. The data for this table were from the union of the joint table of “gene\_pfam” and “pfam\_go”, and the joint table of “gene\_smart” and “smart\_go”. Some manual insertion was also performed. 3053 distinct genes were associated with one or more GO terms.

### 2.3.3 Query and Manipulation Scripts for the *Burkholderia pseudomallei* Genome

#### Database

##### 2.3.3.1 Query scripts

Various query methods are available at <http://origin-sgi.bic.nus.edu.sg/~xiechao/query.html>. The query scripts were written in Perl, and source codes are available to download at <http://origin-sgi.bic.nus.edu.sg/~xiechao/code/>. The database can be queried by gene accession numbers, gene ID, name keywords, gene location, by Pfam/COG domain accession numbers or name keywords, by functional categories, and by Gene Ontology terms.

Most of the query scripts are very short, because they make use of a main, and long, script,

**query.pl** to do the query. Thus they all have the same output format (Figure 2.3).

**Figure 2.3 Sample output of a genome database query (query by name keyword “dnaA”)**

2 genes were found:  
 :::::chromosome 1 : [2 genes](#)  
 :::::chromosome 2 : [0 genes](#)

**Chromosome 1**  
 gene: 1-2

Accession	Position	name	PFam	COG	Functional Class	Gene Ontology	Blast Homolog	Adjacent Genes	Blast	Edit
BP00077	1 : 86871- 85351	probable chromosomal replication initiator protein DnaA <a href="#">[details]</a>	<a href="#">PF00308</a> Bacterial dnaA protein <a href="#">[genes]</a>	<a href="#">COG0593</a> ATPase involved in DNA replication initiation <a href="#">[genes]</a> <a href="#">COG1484</a> DNA replication protein <a href="#">[genes]</a>	Replication, recombination and repair <a href="#">[genes]</a>	<a href="#">GO:0003677</a> DNA binding <a href="#">[genes]</a> <a href="#">GO:0003688</a> DNA replication origin binding <a href="#">[genes]</a> <a href="#">GO:0005524</a> ATP binding <a href="#">[genes]</a> <a href="#">GO:0006270</a> DNA replication initiation <a href="#">[genes]</a> <a href="#">GO:0006275</a> regulation of DNA replication <a href="#">[genes]</a>	PROBABLE CHROMOSOMAL REPLICATION INITIATOR PROTEIN DNAA [Ralstonia solanacearum]	<a href="#">Adjacent Genes</a>	<a href="#">BlastP</a> <a href="#">NCBI Conserved Domain Search</a>	<a href="#">Edit</a>
BP02731	1 : 3371297- 3372022	putative dnaa- related protein <a href="#">[details]</a>	<a href="#">PF00308</a> Bacterial dnaA protein <a href="#">[genes]</a>	<a href="#">COG0593</a> ATPase involved in DNA replication initiation <a href="#">[genes]</a>	Replication, recombination and repair <a href="#">[genes]</a>	<a href="#">GO:0003677</a> DNA binding <a href="#">[genes]</a> <a href="#">GO:0003688</a> DNA replication origin binding <a href="#">[genes]</a> <a href="#">GO:0005524</a> ATP binding <a href="#">[genes]</a> <a href="#">GO:0006270</a> DNA replication initiation <a href="#">[genes]</a> <a href="#">GO:0006275</a> regulation of DNA replication <a href="#">[genes]</a>	CONSERVED HYPOTHETICAL PROTEIN [Ralstonia solanacearum]	<a href="#">Adjacent Genes</a>	<a href="#">BlastP</a> <a href="#">NCBI Conserved Domain Search</a>	<a href="#">Edit</a>

Gene accession number (such as BP01234) is the principle gene extraction method. This is done by script **acc.pl**.

Multiple keywords can be queried for the gene name, and the logic between the keywords can be specified, for instance “ribosomal NOT protein”. This search was achieved by the script **query\_by\_keywords.pl**.

Genes can also be queried by chromosomal position, such as “from position 1000 – 5000

bp at chromosome 1”. This query uses the script **query\_by\_position.pl**.

For Pfam and COG domain query, domain name keywords can be used, or the exact accession number can be used. Scripts **query\_by\_pfam.pl** and **query\_by\_cog.pl** do these queries.

Script **query\_by\_class.pl** can query for the functional categories. All the genes for any functional class can be browsed by this script.

Gene Ontology query is more complicated than others. This is due to the complexity of the GO term tree structure. The script **open\_go.pl** allows user browse the GO tree-structured terms. The tree will also show the gene numbers corresponding each GO term. The corresponded genes to a GO term include not only the directly corresponded genes as shown in the table “gene\_go”, but also the corresponded genes to the term’s child-term, and grandchild-term, etc, as shown in “graph\_path”.

The user who is not familiar with GO terms, or does not like the GO tree structure, can use the script **query\_go\_term.pl** to search GO term by name keywords.

Query by GO term is available for both GO accession number and name keywords search. This is carried out by the script **query\_by\_go.pl**.

For more advanced query, SQL syntax is available by **query\_by\_sql.pl**. However, the user should know the table structure of the database first in order to make this kind of query.

After the output list comes out, more details of each gene can be viewed, for example, the link shown in the “name” column Figure 2.3. The details are showed by the script **show\_detail.pl**.

#### 2.3.3.2 BLAST web server for the *Burkholderia pseudomallei* Genome Database

A BLAST web server for the *B. Pseudomallei* Genome Database was also set up. The address of the server is <http://origin-sgi.bic.nus.edu.sg/xc-bin/blast/blast.html>. The databases on the

BLAST server include both DNA and protein databases. The DNA databases contain either the full length chromosomal sequences or the ORF sequences. The protein database contains the protein sequences of all the genes in the *B. pseudomallei* Genome Database. The protein sequences of the genes for the formatting of the BLAST database were exported into FASTA format by the script **toFasta.pl**.

#### 2.3.3.3 Manipulation scripts

Besides the query scripts, more scripts were written for manipulation of database.

New BLAST and Conserved Domain Database search can be easily achieved by just a single click from either the query output page or the gene detail showing page. The script **blast.pl** gets the protein sequence from the *B. Pseudomallei* Genome Database, and sends it to NCBI BLAST server. The script **cdd.pl** gets the protein sequence from the database and sends it to NCBI CDD search server.

From the new BLAST and CDD search result, if any error in the original database is found, web update interface is also available by two scripts, **update.pl** and **do\_update.pl**. This link can be found from either the query output page or the gene detail page. **update.pl** provides the user the old data for a gene, and collects new data for the gene. **do\_update.pl** does the actual update. It is password protected.

GeneBank format files can be extracted from the database by **togb.pl**. User can specify the location of a chromosomal fragment, and then the **togb.pl** will generate a GeneBank format file that contains all the protein and RNA genes within this chromosomal fragment. The file can be loaded into many biological softwares.

#### 2.3.4 An examples of Applications of the the Database – RND Efflux Identification

*B. pseudomallei* is noted for its intrinsic resistance to many antibiotics (Dance et al. 1989). This is mainly due to its active antibiotics efflux pumps (Moore et al. 1999). The RND family

efflux system, AmrAB-OprA, was reported to be responsible for the efflux of amino glycosides and macrolides in *B. pseudomallei* (Moore et al. 1999). Another two RND family efflux systems have also been characterized in our lab (in publication).

Using the genome database, we have found 9 probable RND multidrug efflux systems (Table 2.2), including the three known. Each of the putative RND transport protein is adjacent to a gene for a probable membrane fusion protein, and most of the RND loci also contain genes for outer membrane proteins of OprM family and a transcriptional regulator.

**Table 2.2 Known and putative RND multidrug efflux pumps in *B. pseudomallei***

Chromosome	Location	Transcriptional regulator	RND membrane fusion protein	RND transporter protein	Outer membrane protein
1	941844-948915	BP00759 <sup>#</sup>	BP00760 <sup>#</sup>	BP00761 <sup>#</sup>	BP00763 <sup>#</sup>
1	324453-330545		BP00302	BP00303	BP00300
1	1462488-1471419	BP01200	BP01199	BP01197, BP01198	
1	1817818-1824836	BP01513	BP01512	BP01511	BP01509
1	2146271-2152980	BP01730 AAC27752*	BP01729 AAC27753*	BP01728 AAC27754*	BP01727
1	2690327-2694717		BP02159	BP02160	
2	388854-396971	BP03611 <sup>#</sup>	BP03613 <sup>#</sup>	BP03614 <sup>#</sup>	BP03615 <sup>#</sup>
2	1416409-1423063	BP04355	BP04357	BP04356	BP04358
2	1498963-1505662	BP04434	BP04435	BP04436	BP04437

\*GeneBank accession number of the AmrAB-OprA RND efflux systems (Moore et al. 1999)

<sup>#</sup> RND efflux systems that have been characterized in our lab.

*B. pseudomallei* contains many more predicted RND efflux systems than *E. coli* (4 pumps), *B. subtilis* (1 pump), and *M. tuberculosis* (none), but similar with *P. aeruginosa* (10 pumps), which is another organism resistant to many antibiotics (Stover *et al.* 2000).

## **2.4 Discussion**

The *Burkholderia pseudomallei* genome was annotated at both protein-level and functional class- or process-level. The annotation was stored in the *B. pseudomallei* Genome Database. Web interface of the database was also developed, therefore further

analysis of the *B. pseudomallei* genome will be much more easily.

Although we have put our best effort to make the genome database more reliable, there could be errors in the database. The errors can be introduced in several ways. For example, in the protein-level annotation, the gene names were annotated by choosing from the names of its homologs, however, the homolog names from GeneBank may not be reliable.

Manual check of every gene annotation is required, though human may also introduce errors. Experts in various research fields may help reduce the error probability by checking and confirming the annotation results by programs. However, this requires lots of time and human resources. Actually, the sequencing of *B. pseudomallei* was finished in May 2003 in the Sanger Institute, but their annotation by large group still has not been out yet.

Although errors can not be avoided, the error rate is low enough for most genome scale analysis. The Pfam, Smart, and COG domain annotation for the genes were directly extracted from NCBI CDD searches; therefore, there is little chance to introduce errors. The functional categories annotation was bridged by COG, and therefore the reliability of this annotation is as high as the reliability of COG-functional class mapping. However, the Gene Ontology annotation is not as reliable as the other parts. One of the reasons is that the reliability of mapping files from Pfam/Smart to GO terms is not very high. Another reason is that even the GO vocabulary is not complete. New version of GO vocabulary is released every month. Thus, periodically update of the database is also required.

Despite the possibilities of errors in the *B.pseudomallei* Genome database, it is very helpful for researchers. It was made available online, any researcher in the field of *Burkholderia* can now access the annotation information and accelerate their pace of research. Take the RND efflux systems as an example; we can easily identify 9 RND systems in the *B. pseudomallei* genome with the help of the genome database.

# CHAPTER 3. ANALYSES OF THE *BURKHOLDERIA PSEUDOMALLEI* GENOME

## **3.1 Introduction**

*Burkholderia pseudomallei* belongs to beta-proteobacteria, a group of bacteria whose genomic organization is still poorly characterized. When compared to other bacteria, it has a very large genome, which comprises two replicons. Both replicons are recognized as chromosomes by researchers involved in its sequencing (<http://www.sanger.ac.uk>). This is very different with most of the other bacterial genomes, which have only one chromosome usually.

In this chapter, the *B. pseudomallei* genome was analyzed using the genome database constructed in the previous chapter. The first part of the study investigates the probable evolution history of its large genome. This provides a clue as to whether its large genome arose from recent gene duplication or functional diversity.

Secondly, many features of the genome, including origin of replication, functional categories, pathways, transcriptional regulators, are studied in totality for the whole genome as well as for two chromosomes separately. The results were used to illustrate the differences between the two replicons and to predict if any of the two replicons may be extrachromosomal or dispensable.

## **3.2 Material and Methods**

### **3.2.1 Materials**

The analysis was based on the *Burkholderia pseudomallei* Genome Database that was constructed in Chapter 2. The database contains the annotated genome of strain K96243. The 20 Feb 2004 version of this database was used. The backup file of this version can be found

at <http://origin-sgi.bic.nus.edu.sg/~xiechao/data/>. Most BLAST (Altschul *et al.* 1997) searches were performed on local machine. The BLAST program of version 2.2.6 were downloaded from the NCBI ftp server.

Both Java and Perl were used for programming. All the source codes of the programs written in this project are available at <http://origin-sgi.bic.nus.edu.sg/~xiechao/code/>.

### 3.2.2 Comparative genomics

The proteome of *B. pseudomallei* was compared with the proteomes of all the microbial proteomes that have been completely sequenced (155 genomes up to 25 Feb 2004), both separately and together. The comparison was using the program **blastp**. The hits with expect value of less than  $1e-5$  for each proteome were recognized as homologous. The protein sequences of the 155 genomes were downloaded from NCBI ftp server. Perl scripts and Java programs were written to do the BLAST search and analyze of the BLAST results. Programs include **splitfasta.pl**, **format.pl**, and **blaster.pl** for BLAST search; and **extractCompare.java**, **extractCompare2.java**, and **countGroup.java** for analysis.

### 3.2.3 Prediction of paralogous groups

Paralogous groups were constructed by comparing predicted ORFs using BLAST. Each ORF was used to do **blastp** against the database of the whole proteome. A Perl script **blaster.pl** was used to carry out the BLAST. Then the hits containing a minimum of 50% identity over 60% of the length of the smaller ORF were clustered into families. Two Java programs **readBlast.java** and **analyzeParalog.java** were used to analyze the BLAST results.

### 3.2.4 Prediction of origin of replication

The origin of replication was predicted by GC skew analysis, based on the bias towards G on the leading strand during replication (Lobry 1996). The difference of the frequencies of C and G was calculated as  $(C-G)/(C+G)$ . The origin of replication was indicated by the switch

of (C-G)/(C+G) from positive to negative. The calculation was using a window size of 20 kbp and a sliding step of 5 Kbp. A Java program **origin.java** was written to do the calculation. The direct repeats of the binding sites of the replication initiator protein were calculated using the Java program **drepeat.java**.

### **3.2.5 Functional categories**

The results of functional categories were summarized from the *B. pseudomallei* Genome Database. The classification in the database was based on the COG domains (Tatusov *et al.* 1997).

### **3.2.6 Pathway prediction**

The pathway prediction was based on the COG domain search results. A Java program **pathwayCOG.java** was written to extract the COG-pathway relationships from NCBI COG pathway pages (<http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?sys=all>). Then a Perl script **pathwayGene.pl** was written to search the genes with a particular COG domain from the *B. pseudomallei* Genome Database for each pathway.

### **3.2.7 Potential transcriptional regulators prediction**

The transcriptional regulators were predicted using the Pfam domain searches. The transcriptional regulator domains were extracted from the Pfam database (Bateman *et al.* 2002). An ORF containing the domains was classified as a transcriptional regulator if the Expect value is less than 1e-4.

### **3.2.8 Statistical analysis**

We performed statistical analysis using Fisher's exact test (Fisher 1922). P-values less than 0.05 were considered as significant.

### **3.3 Results**

#### **3.3.1 General features of the genome**

The general features of *B. pseudomallei* genome were shown in Table 3.1. It has two replicons: a larger chromosome of 4074542 bp and a smaller chromosome of 3173005 bp, yielding a total genome size of 7247547 bp. This is larger than most of the 155 completely sequenced bacteria genomes (up to 25 Feb 2004). The two chromosomes have a similar G+C ratio, and almost identical protein coding percentage. However, there are still some differences between the two chromosomes. The average length of proteins of chromosome 1 is much smaller than that of chromosome 2, and most of the tRNAs and rRNAs are in chromosome 1.

**Table 3.1 General Features of the *B. pseudomallei* genome**

	Chromosome 1	Chromosome 2	Genome
Length (bp)	4074542	3173005	7247547
G+C ratio	67.7%	68.5%	68.1%
Protein-coding region	78.1%	77.9%	78.0%
Protein-coding genes	3327	2352	5679
Average length of proteins	319	351	332
tRNA	52	7	59
Ribosomal RNA operons	3	1	4

#### **3.3.2 Paralogous groups in *B. pseudomallei***

In order to investigate whether the large genome of *B. pseudomallei* was due to recent gene duplication or more functional diversity, the gene families (paralogous groups) of *B. pseudomallei* was predicted by **blastp** against itself. If the large genome of *B. pseudomallei* was arisen by recent gene duplication, the number of paralogous groups should be similar to, while the average sizes of the paralogous groups should be larger than, the other bacteria, such as *E. coli*, *B. subtilis*, and *M. tuberculosis*. The paralogous grouping results are shown in Table 3.2.

**Table 3.2 Paralogous groups in *B. pseudomallei***

	Chromosome 1	Chromosome 2	Genome	<i>E. coli</i> *	<i>B. subtilis</i> *	<i>M. tuberculosis</i> *
No. of groups	74	98	253	166	166	185
% of ORFs in groups	4%	9%	10%	9%	9%	13%
Mean group size	2.2	2.1	2.2	2.4	2.2	2.7

\* These results were taken from Stover *et al.* (2000)

There are 253 paralogous groups in *B. pseudomallei*. The number of paralogous groups in *B. pseudomallei* was much higher than the other 3 genomes listed in the table; while, the average group size is similar. These data indicate that the *B. pseudomallei* genome may have evolved by incorporating several small paralogous gene families, whose members encode distinct functions, not by simple duplication and therefore there is no increase of average group size of the existent paralogous gene families. It is also noted that there are only 4% genes are in paralogous groups in chromosome 1, less than half of that in chromosome 2. This shows that the genes in chromosome 1 are functionally more distinct than in chromosome 2 and, due to its larger size, may also encode more complete sets of proteins with diverse functions. This will be proved by the following sections.

### 3.3.2 Comparative genomics

The *B. pseudomallei* proteome was firstly BLAST against each of the other bacteria that have been completely sequenced, in order to find the genetically most similar organism. The results are shown in Table 3.3.

**Table 3.3 The top 10 organisms with best similarity with *B. pseudomallei* genome**

Accession number	Organism	ORFs in organism database	ORFs in <i>B. pseudomallei</i>	Hits with E<1e-5	% of hits over <i>B. pseudomallei</i> ORF
NC_003295 NC_003296	<i>Ralstonia solanacearum</i> with its Megaplasmid	5116	5658	3969	70.1%
NC_002516	<i>Pseudomonas aeruginosa</i>	5567	5658	3708	65.5%
NC_005085	<i>Chromobacterium violaceum</i>	4407	5658	3544	62.6%
NC_003295	<i>Ralstonia solanacearum</i>	3440	5658	3497	61.8%
NC_002927	<i>Bordetella bronchiseptica</i>	4994	5658	3495	61.7%
NC_002947	<i>Pseudomonas putida</i>	5350	5658	3486	61.6%
NC_004463	<i>Bradyrhizobium japonicum</i>	8317	5658	3437	60.7%
NC_004578	<i>Pseudomonas syringae</i>	5471	5658	3424	60.5%
NC_002928	<i>Bordetella parapertussis</i>	4185	5658	3410	60.2%
NC_002678	<i>Mesorhizobium loti</i>	6746	5658	3273	57.8%

*B. pseudomallei* is most similar with *Ralstonia solanacearum*. They are both members of *Burkholderiaceae* (Salanoubat *et al.* 2002). 70.1% of *B. pseudomallei* ORFs have homologs in *R. solanacearum* proteome. The second most similar organism is *P. aeruginosa*, and the third one is *C. violaceum*, which are gamma- and beta-division of proteobacteria respectively (Ribeiro De Vasconcelos *et al.* 2003; Stover *et al.* 2000).

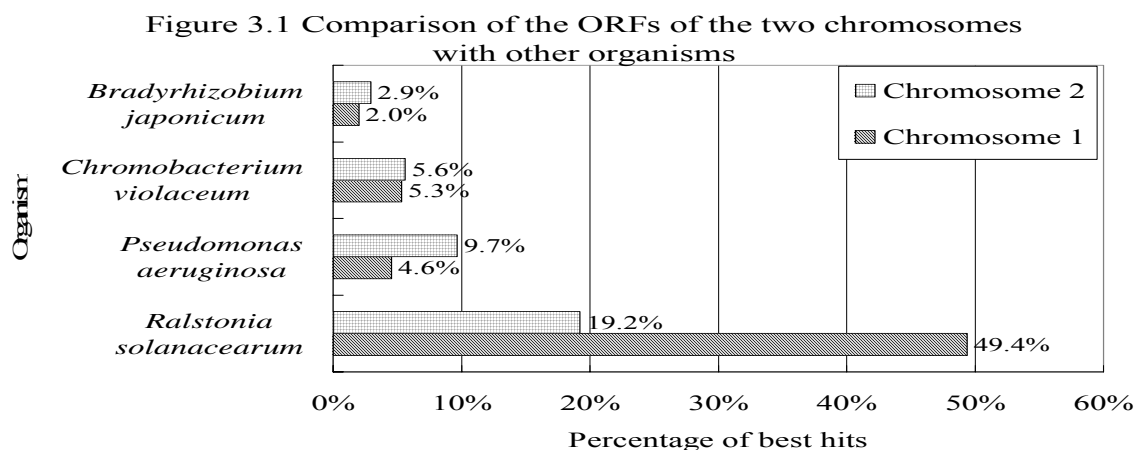
In order to test whether the ORFs on the two chromosomes have different origins, each of the predicted ORFs was used as the query sequence for a **blastp** query against a pool of the ORFs of the 155 fully sequenced bacterial genomes. The results are shown in Table 3.4.

49.4% of the best hits of chromosome 1 are from *R. solanacearum*. The next largest groups of the best hits of chromosome 1 are *C. violaceum* (5.3%) and *P. aeruginosa* (4.6%). The best hits of chromosome 2 are more distributed than chromosome 1. The largest source of the best hits is also *R. solanacearum*, but only 19.2% of the chromosome 2 ORFs. The next largest groups of the best hits of chromosome 2 are *P. aeruginosa* (9.7%) and *C. violaceum* (5.6%).

**Table 3.4 Comparison of *B. pseudomallei* ORFs on the two chromosomes with other organisms.** The best hits of the ORFs were categorized as their source organisms. Both number and percentage of best hits of each chromosome and the whole genome were shown.

Organism	genome		Chromosome 1		Chromosome 2		p-value
	Num	%	Num	%	Num	%	
<i>Ralstonia solanacearum</i>	2095	36.9%	1643	49.4%	452	19.2%	0.000
<i>Pseudomonas aeruginosa</i>	379	6.7%	152	4.6%	227	9.7%	0.000
<i>Chromobacterium violaceum</i>	309	5.4%	177	5.3%	132	5.6%	0.635
<i>Bradyrhizobium japonicum</i>	136	2.4%	67	2.0%	69	2.9%	0.028
<i>Bordetella pertussis</i>	128	2.3%	82	2.5%	46	2.0%	0.237
<i>Pseudomonas putida</i>	124	2.2%	61	1.8%	63	2.7%	0.034
<i>Pseudomonas syringae</i>	123	2.2%	58	1.7%	65	2.8%	0.012
<i>Bordetella parapertussis</i>	106	1.9%	55	1.7%	51	2.2%	0.164

The distribution over the two chromosomes of the best hits from *R. solanacearum* is strongly biased ( $p = 1.4e-124$ ). As shown in Figure 3.1, 1643 of the hits are on chromosome 1 (49.4% of chromosome 1 ORFs), while only 452 hits are on chromosome 2 (19.2% of chromosome 2 ORFs). The distribution of best hits from *P. aeruginosa* is also biased ( $p = 7.9e-14$ ) (Figure 3.1). The hits are favor chromosome 2 (227 hits, 9.7%) over chromosome 1 (152 hits, 4.9%).



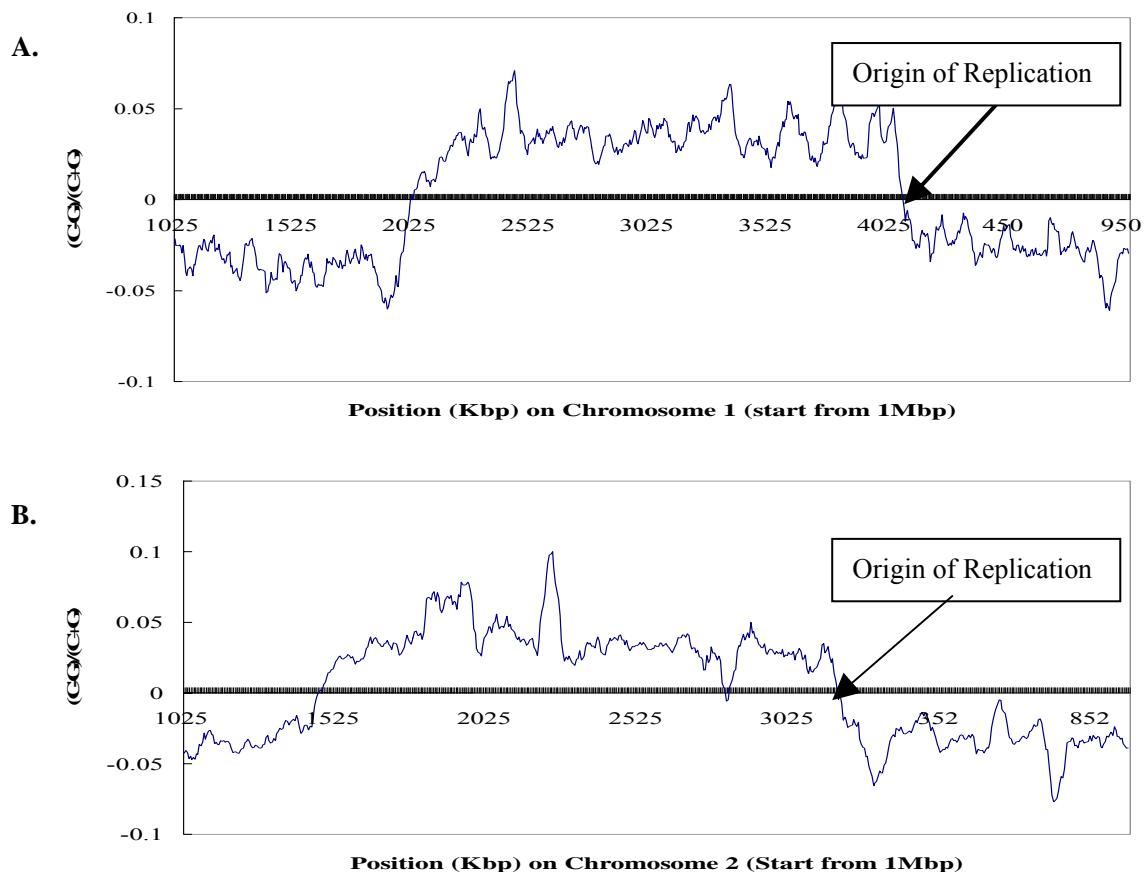
These data indicate that chromosome 1 is more close to *R. solanacearum*, with 49.4% of ORFs have best hits from *R. solanacearum*. However, there is no dominant source of best hits for chromosome 2, The largest group of the best hits for chromosome 2 are only 19.2% of the ORFs. The distribution of best hits from *P. aruginosa* is strongly favored to chromosome 2. Therefore, we propose that chromosome 2 is genetically more close to *Pseudomonas*. Supporting this hypothesis, the distribution of best hits from another two *Pseudomonas* species, *P. putida* and *P. syringae*, are also favored to chromosome 2 ( $p = 0.034$ ,  $p = 0.012$ ); while the distribution best hits of other beta-proteobacteria are not biased (*C. violaceum*,  $p = 0.635$ ; *B. pertussis*  $p = 0.237$ ; *B. parapertussis*  $p = 0.164$ ). Actually *Burkholderia pseudomallei* was classified as *Pseudomonas pseudomallei* previously (Sexton *et al.* 1993). To explain that the largest source of best hits for chromosome 2 is also *R. solanacearum*, we propose that DNA recombination during evolution have mixed the genes on chromosome 1 and chromosome 2.

### 3.3.3 Origin of Replication in *B. pseudomallei*

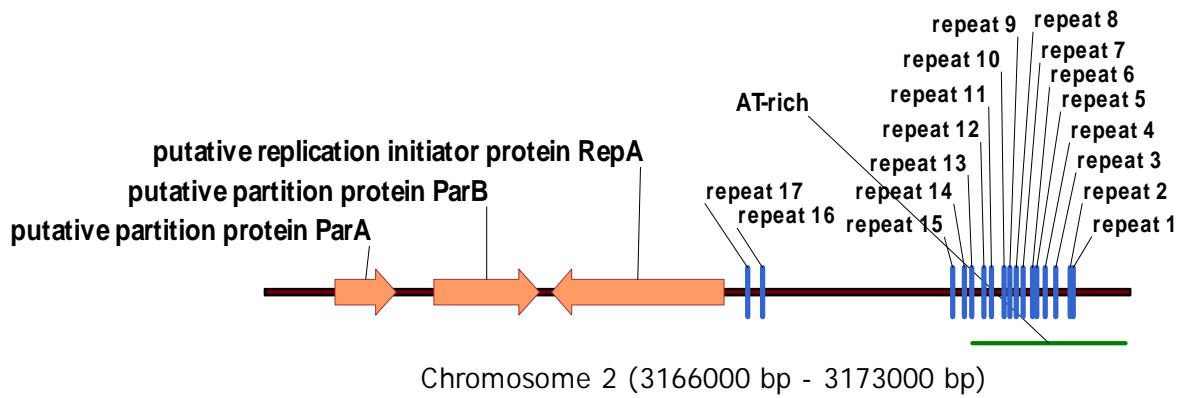
The origins of replication of both chromosomes were predicted by the GC skew analysis, based on the bias towards G on the leading strand during replication. The results were shown in Figure 3.2. The origin of replication of chromosome 1 was around 15kbp (Figure 3.2 A). The replication initiator protein for chromosome 1 is the putative DnaA protein (Accession number of the *B. pseudomallei* Genome Database: BP00077), located from 85351 to 86871 bp at complement strand of chromosome 1. The origin of replication of chromosome 2 was around 3170kbp (Figure 3.1 B). This origin of replication has characteristics of plasmid borne *ori* loci. It was shown in Figure 3.2. It is flanked by a putative plasmid replication initiator protein RepA (BP05679). This gene contains the Pfam domain PF01051 (initiator of plasmid replication. Table 3.5). There are also 15 repetitions of a 17bp conserved motif, (consensus

sequence: CCCG[AT]AAAA[TC]CTCACCT). They are putative RepA-binding boxes, or iterons, and located in an AT-rich region (51% of G+C over 1200bp). Two more repeats of the putative RepA-binding boxes are around the promoter region of RepA. They may function as auto-repressor of the RepA expression and thus control the replication cycle. Two putative plasmid active partitioning protein ParA (BP05678) and ParB (BP05679) are located downstream of the putative RepA (Figure 3.3, Table 3.5). From these results, the chromosome 2 appears to be a megaplasmid.

**Figure 3.2 The GC skew analysis of Chromosome 1 (A) and 2 (B).** The start position for each chromosome is the 1 millionth base. The putative origin of replication was indicated by the arrows. (Window size = 20kbp, sliding step = 5kbp)



**Figure 3.3 The origin of replication and the replication proteins of chromosome 2.**



**Table 3.5 The replication proteins on chromosome 2.**

Pfam domains and the scores were showed (S, score; E, expect)

Accession	Name	Pfam Domain	Score
BP05677	Putative partition protein ParA	PF00991 ParA family ATPase	S = 44.9 E = 7e-6
BP05678	Putative partition protein ParB	PF02195 ParB-like nuclease domain	S = 53.8 E = 4e-8
BP05679	Putative replication initiator protein RepA	PF01051 Initiator replication protein	S = 51.9 E = 3e-7

### 3.3.4 Functional categories

In order to investigate the different functional status of the two chromosomes in *B. pseudomallei* genome, the predicted genes were classified into functional categories based on COG domains. 62.8% of ORFs were assigned one or more functional class (Table 3.6); however, the distribution of these ORFs is significantly biased in favour of chromosome 1 ( $p = 2.5e-9$ ). In other words, there are significantly more genes in chromosome 2 that can not be classified into any of the functional categories, due to the lack of COG domains matches.

**Table 3.6 Functional categories of predicted ORFs**

	Chromosome 1	Chromosome 2	P-value
Translation, ribosomal structure and biogenesis	206	35	0.000*
Transcription regulators	266	241	0.004*
DNA replication, recombination and repair	131	39	0.000*
Cell division and chromosome partitioning	47	18	0.031*
Posttranslational modification, protein turnover, chaperones	179	84	0.001*
Cell envelope biogenesis, outer membrane	286	170	0.067
Cell motility and secretion	110	77	1.000
Inorganic ion transport and metabolism	280	184	0.432
Signal transduction mechanisms	158	152	0.006*
Energy production and conversion	251	202	0.164
Carbohydrate transport and metabolism	295	179	0.098
Amino acid transport and metabolism	464	293	0.104
Nucleotide transport and metabolism	126	60	0.010*
Coenzyme metabolism	267	116	0.000*
Lipid metabolism	173	146	0.114
Secondary metabolites biosynthesis, transport and catabolism	151	165	0.000*
Total distinct genes with a functional category	2189	1364	0.000*

\*Functional categories that are biased in distribution over the two chromosomes ( $p < 0.05$ )

The distribution of some functional categories between chromosome 1 and chromosome 2 are significantly biased. Categories that distribute biased in favour of chromosome 1 include the class of “DNA replication, recombination and repair” ( $p = 3.3e-7$ ), “cell division and chromosome partitioning” ( $p = 0.031$ ), “translation, ribosomal structure and biogenesis” ( $p = 3.6e-20$ ), “nucleotide transport and metabolism” ( $p = 0.010$ ), “posttranslational modification, protein turnover, chaperones” ( $p = 0.001$ ), and “coenzyme metabolism” ( $p = 3.6e-6$ ). Most of these classes are essential housekeeping genes. Therefore, it appears that chromosome 1 is more essential than chromosome 2, in the terms of bacteria proliferation.

There are also a few functional categories that distribute significant biased in favour of chromosome 2, including the class of “transcription regulators” ( $p = 0.004$ ), “signal transduction mechanisms” ( $p = 0.006$ ), and “secondary metabolites biosynthesis, transport, and catabolism” ( $p = 7.8e-5$ ). These categories are involved in more advanced cellular functions, transcriptional control and secondary metabolites metabolism.

From these data, we can see that chromosome 1 is involved in more basic housekeeping functions, while chromosome 2 is involved in more advanced functions, such as transcriptional control and signal transduction.

Now, we want to ask whether chromosome 1 carries a complete set of the housekeeping genes. Pathways identification of *B. pseudomallei* may help us answer this question. This analysis is carried out in section 3.3.5. Another thing we need pay attention to is the fact that transcriptional regulators and signal transduction systems biased on chromosome 2, therefore, we may find out the functional status of chromosome 2 in the genome by comparing these genes of the two chromosomes with other bacteria. This will be done in section 3.3.6.

### **3.3.5 Pathways Identification in *B. pseudomallei***

As we have known in section 3.3.4, chromosome 1 carries significantly more housekeeping genes than chromosome 2. In order to investigate whether chromosome 1 carry a complete set of the essential genes for bacteria to survive, the pathways were identified based on COG domains.

The results were shown in Table 3.7 – 3.13. In these tables, total COG numbers of each complete pathway, and the number of COGs found were shown. (For details of the pathway identification please see the supplementary data at <http://origin-sgi.bic.nus.edu.sg/~xiechao/data/>). In the following analysis, the threshold for the completeness of a pathway was defined here as more than 70% of the COGs of a pathway can be found.

The amino acid biosynthesis pathways were listed in the Table 3.7. Genes in chromosome 1 can independently carry out all the amino acid biosynthesis pathways. However, chromosome 2 encodes only the complete set of enzymes for leucine biosynthesis pathway, although it carries some incomplete set of genes in other amino acid biosynthesis pathways.

**Table 3.7 Amino Acid Biosynthesis pathways in *B. pseudomallei*.**

Pathway	Total COGs in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
Phenylalanine/Tyrosine Biosynthesis	14	13	7	13
Arginine Biosynthesis	11	11	6	11
Threonine Biosynthesis	5	5	2	5
Tryptophan Biosynthesis	17	12	9	16
Isoleucine Biosynthesis	6	6	4	6
Valine Biosynthesis	6	6	4	6
Leucine Biosynthesis	10	8	8	10
Methionine Biosynthesis	10	9	4	9
Proline Biosynthesis	5	4	3	5
Histidine Biosynthesis	12	11	3	12

The nucleotide biosynthesis and salvage pathways were shown in Table 3.5. Chromosome 1 encodes all the enzymes necessary for nucleotide biosynthesis and salvage whereas chromosome 2 does not.

**Table 3.8 Nucleotide biosynthesis/salvage pathways in *B. pseudomallei*.**

Pathway	Total COG in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
Purine Salvage	5	4	1	4
Purine Biosynthesis	18	17	7	17
Pyrimidine Salvage	10	7	5	8
Pyrimidine Biosynthesis	14	14	4	14
Thymidylate Biosynthesis	10	9	2	9

The pathways in energy production and conversion were similarly analyzed and shown in

Table 3.9. Chromosome 1 has complete sets of genes in 7 of the 9 pathways, while chromosome 2 has only 2 sets. According to the COG domain search results, chromosome 1 does not encode all the genes necessary for glycolysis and pyruvate decarboxylation. Chromosome 2, however, encodes the complete set of enzymes of the TCA cycle and fatty acid biosynthesis pathways.

**Table 3.9 Energy production and conversion pathways in *B. pseudomallei*.**

Pathway	Total COG in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
TCA Cycle	16	13	14	16
Glycolysis	14	9	1	9
Entner-Doudoroff Pathway	4	4	0	4
Gluconeogenesis	14	10	0	10
Pyruvate Decarboxylation	7	4	4	6
Lipid A Biosynthesis	9	9	2	9
Glyoxylate Bypass	2	2	1	2
Pentose Phosphate Pathway	9	8	4	9
Fatty Acid Biosynthesis	13	13	12	13

Chromosome 1 carries the complete sets of genes for the electron transport and oxidative phosphorylation pathways, while chromosome 2 has only one of the three pathways (Table 3.10).

**Table 3.10 Electron transport and oxidative phosphorylation in *B. pseudomallei*.**

Pathway	Total COG in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
Ubiquinone Biosynthesis	15	14	8	14
F0f1-Type Atp Synthase Subunits	9	7	7	8
Nadh:Ubiquinone Oxidoreductase Subunits	15	14	7	14

Chromosome 1 has complete sets of genes for all the 10 pathways involved in coenzyme biosynthesis, while chromosome 2 has none (Table 3.11).

**Table 3.11 Coenzyme biosynthesis pathways in *B. pseudomallei*.**

Pathway	Total COG in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
Nad Biosynthesis	7	6	2	6
Fad Biosynthesis	9	8	2	8
Heme Biosynthesis	14	13	4	13
Biotin Biosynthesis	6	5	2	5
Thiamine Biosynthesis	10	9	4	9
Riboflavin Biosynthesis	7	6	2	6
Pyridoxal Phosphate Biosynthesis	8	7	3	7
Coenzyme A Biosynthesis	9	8	2	8
Cobalamin Biosynthesis	18	17	5	17
Menaquinone Biosynthesis	16	14	9	14

. Chromosome 1 carries complete sets of genes for all the replication, transcription and translation basal machineries, whereas chromosome 2 has none complete (Table 3.12).

**Table 3.12 Replication/Transcription/Translation basal machinery in *B. pseudomallei*.**

Pathway	Total COG in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
Basal Replication Machinery	24*	19	5	21
Ribosomal Proteins - Small Subunit	21*	20	0	20
Ribosomal Proteins - Large Subunit	31*	30	0	30
DNA Polymerase III Subunits	8	7	0	7
DNA-Dependent RNA Polymerase Subunits	8*	7	4	9
Aminoacyl-tRNA Synthetases And Alternative Systems For Amino Acid Activation	26	24	8	24
Translation Factors And Enzymes Involved In Translation	17*	14	3	14

\* The number was modified to suit for prokaryotes.

Some other pathways were shown in Table 3.13. Chromosome 1 has complete sets of genes for all the three pathways, but chromosome 2 has none.

**Table 3.13 Other pathway or machinery in *B. pseudomallei*.**

Pathway	Total COG in the pathway	COGs found in Chromosome 1	COGs found in Chromosome 2	COGs found in Genome
Flagellum Structure And Biogenesis	33	30	12	30
Preprotein Translocase Subunits	9	8	2	8
Deoxyxylulose Pathway Of Terpenoid Biosynthesis	5	5	1	5

All the results (Table 3.7 – 3.13) show that chromosome 1 encodes a complete set of essential housekeeping proteins including those required for (1) DNA replication, cell division; (2) transcription; and (3) translation. This last set includes genes encoding all the ribosomal proteins, 3 complete ribosomal DNA loci, and 52 transfer RNAs. All the essential genes required for the purine and pyrimidine biosynthesis and salvage can be found on chromosome 1. None of them is complete on chromosome 2. Finally, all the amino acid biosynthesis, coenzyme biosynthesis, electron transport and phosphorylation, and most of the energy production and conversion pathways can be completely found on chromosome 1, while very few of them can be found on chromosome 2.

Chromosomal 2 also carries some metabolically essential genes; however, they are also present on the chromosome 1. These include genes of many incomplete pathways and a complete ribosomal DNA locus with 2 tRNAs, which is identical to the 3 in chromosome 1 and possibly derived from duplication of rDNA loci on chromosome 1.

Together with the origin of replication identification results, we propose that since chromosome 1 encodes all of the basal genes required for the survival of the bacterium, it should be referred to as the “chromosome” whereas chromosome 2 could be referred to as “megaplasmid”.

### **3.3.6 Transcriptional regulators in *B. pseudomallei***

As shown in section 3.3.4, the distribution of transcriptional regulators and signal

transduction systems is significantly biased in the favour of chromosome 2 based on COG domains and functional categories. In this part, this conclusion is confirmed by Pfam domain searches, and more details were shown. The portion of transcriptional regulators over the proteome of each chromosome was compared with other bacteria, in order to investigate the function of chromosome 2 in the genome. The results were shown in Table 3.14.

**Table 3.14 Transcriptional regulator domains in *B. pseudomallei***

Regulator motif	Pfam Accession	Chromosome 1	Chromosome 2	Genome
AraC	PF00165 PF02311	15	30	45
Crp	PF00325	0	3	3
GntR	PF00392	10	15	25
LysR	PF00126 PF03466	44	50	94
Cold shock	PF00313	1	1	2
LacI	PF00356	4	0	4
AsnC	PF01037	10	6	16
MerR	PF00376	6	1	7
LuxR	PF00196	9	16	25
Sigma54	PF00158	8	6	14
Sigma70	PF04542 etc.	11	6	17
ArsR	PF01022	1	4	5
DeoR	PF00455	2	2	4
IclR	PF01614	5	8	13
TetR	PF00440	9	12	21
MarR	PF01047	8	6	14
SIS domain	PF1380	11	2	13
Two-component systems				
Sensors	PF00512	17	19	36
Response regulators	PF00072	28	33	61
Hybrids*	*	4	5	9
Total				
Total		199	220	446
Distinct		188	209	397
% predicted regulator		5.7%	8.9%	7.0%

\* Hybrids were ORFs which contain both response regulator and sensor motifs.

The data shows that 397 genes containing motifs characteristic of transcriptional regulators or environmental sensors. These regulator genes are favored on chromosome 2, 188 genes on chromosome 1 and 209 genes on chromosome 2 ( $p = 3.1 \times 10^{-6}$ ). Therefore, it confirmed the conclusion from section 3.3.4.

This analysis predicts that 7.0% of *B. pseudomallei* genes are involved in transcriptional regulation or signal transduction. This ratio is higher than most other completely sequenced genomes (*E. coli* 5.8%, *B. subtilis* 5.3%, *M. tuberculosis* 3.0%, *H. pylori* 1.1%), but lower than *P. aeruginosa* (8.4%) (Stover *et al.* 2000). This ratio on chromosome 1 is 5.7%, which is similar with the ratio in *E. coli* and *B. subtilis*, while the ratio on chromosome 2 is 8.9%, even higher than the ratio in *P. aeruginosa*.

From these results, we can predict that the high portion of transcriptional regulators in *B. pseudomallei* was due to the chromosome 2. Without chromosome 2, *B. pseudomallei* may be similar to *E. coli*, or *B. subtilis*, while the presence of chromosome 2 makes the genome of *B. pseudomallei* more complicated and robust, like that of *P. aeruginosa*, another pathogen with complicated genome structure. With the advanced functions of genes on chromosome 2, *B. pseudomallei* can grow in diverse range of environments, from soil, plant, to animal, and become a pathogen for both plant and animal.

### **3.4 Discussion**

The genome of *Burkholderia pseudomallei*, previously known as *Pseudomonas pseudomallei*, contains 7.2 Mbp, which is larger than most of the other completely sequenced genomes. Paralogous grouping analysis showed that the larger genome may have arisen from functional diversity, instead of recent gene duplication.

Comparative genomics showed that *B. pseudomallei* is most similar with *R. solanacearum*, another member of *Burkholderia*aeas, beta-proteobacteria. The second most

similar organism is *P. aeruginosa*, a gamma-division of proteobacteria. Further analysis showed that genes on chromosome 1 share more pure source than that of chromosome 2. Furthermore, *Pseudomonaeas*-like genes distribute significantly biased in favour of chromosome 2. Therefore, we propose that chromosome 2 is the source of the functional complexity of *B. pseudomallei* genome.

There are two replicons in *B. pseudomallei* genome, and both were recognized as “chromosome” (<http://www.sanger.ac.uk/pathogen/bpm/>). They have similar G+C ratio, similar protein coding percentage. However, in this study, we found some evidences that showed chromosome 2 is significantly different from chromosome 1, such as origin of replication and gene function differences of the two replicons. These differences indicate that the chromosome 2 is probably a dispensable genetic element, and should be called a megaplasmid.

From the GC skew analysis, chromosome 2 has a plasmid type origin of replication. This origin of replication is flanked by the putative plasmid replication initiator protein RepA, and 15 direct 17bp-repeats, which are probable RepA binding boxes, or iteron. These are characteristics of plasmid origin of replication (del Solar *et al.* 1998). Putative plasmid partitioning protein ParA and ParB are also downstream of the putative RepA. These data strongly suggest that chromosome 2 uses the theta mechanism of plasmids to do replication, and therefore, it could be a megaplasmid, not chromosome.

Further analysis of the gene contents of chromosome 1 and 2 showed that the larger replicon (chromosome 1) contains significantly more housekeeping genes than the smaller one, while the smaller replicon contains more transcriptional regulators and signal transduction systems. Upon further analysis, the larger replicon was found to carry all the genes of essential pathways, thus suggesting that the smaller replicon might be dispensable.

Since the smaller replicon also contains some genes belonging to the essential biochemical pathways, we propose that their function is to improve the fitness of the *B. pseudomallei* for survival in very diverse environment. In other words, without the smaller replicon, the bacterium's survival might be restricted to only a specific niche, e.g. in soil or water but not in animals and humans. Supporting this hypothesis is the finding that significantly more transcriptional regulators and signal transduction systems are encoded by the smaller replicon, the functions of which may be useful for the regulation of gene expressions in diverse environments.

Experimental approach can be carried out to check whether chromosome 2 is dispensable megaplasmid. Loss of chromosome 2 in natural strains of the *B. pseudomallei* can be screened. If there are natural strains that can survive without chromosome 2, it proves that chromosome 2 is a dispensable megaplasmid. Alternatively, we could attempt to knock out the replication of chromosome 2 or to cure the cells of the megaplasmid. If the *B. pseudomallei* without is viable, then we can conclude that the chromosome 2 is dispensable.

# **CHAPTER 4. DETECTION OF PATHOGENICITY ISLANDS AND PUTATIVE ALIEN GENE CLUSTERS IN THE *BURKHOLDERIA PSEUDOMALLEI***

## **4.1 Introduction**

*Burkholderia pseudomallei*, a Gram-negative saprophytic soil bacterium, is also the aetiological agent of melioidosis, a potentially fatal infectious disease occurring in man and animal (Leelarasamee and Bovornkitti 1989). Pathogenesis of melioidosis is poorly understood. *B. pseudomallei* has a broad range of virulence factors that are likely to influence pathogenesis and to be responsible for the various clinical presentations of melioidosis. It can secrete substances that cause tissue necrosis, haemolytic cytolysis and death (Ashdown and Koehler 1990; Leelarasamee and Bovornkitti 1989; Sexton *et al.* 1994).

Most of the recent research has been focused on the identification of virulence factors of *B. pseudomallei* and the development of diagnostic methods to melioidosis. These virulence factors are often encoded within clusters of multiple genes that are involved in the process of pathogenesis. Such gene clusters are called pathogenicity islands (PAIs) (Hacker and Kaper 2000). PAIs are present in the genomes of pathogenic organisms but absent from the genomes of nonpathogenic organisms of the same or closely related species, and they contain genes involved in diseases, such as genes encoding invasions, adhesions and secretion factors, and are often sources of toxins. They often consist of DNA regions that differ from the whole genome in G+C content and in codon usage, which may reflect the generation of PAIs by horizontal gene transfer. PAIs are often flanked by small directly repeated (DR) sequences. These sequences may be generated after integration of PAI-specific DNA regions into the host genome via recombination. PAIs are often associated with transfer RNA genes. tRNA

loci often act as integration sites for foreign DNA. The association of PAIs and tRNA loci may therefore reflect the generation of PAIs by horizontal gene transfer. PAIs often carry cryptic or functional genes encoding mobility factors such as integrase, transposases, and insertion sequence (IS) elements or parts of these elements. PAIs often do not represent homogeneous pieces of DNA but rather are made up of mosaic-like structures which have been generated by a multi-step process. They often represent unstable DNA regions, whose deletions may occur via the direct repeats (DRs) at their ends or via IS elements or other homologous sequences located on PAIs (Hacker and Kaper 2000).

In *B. pseudomallei* and many other Gram-negative bacteria, PAIs have been associated with type III secretion systems (Attree and Attree 2001; Hueck 1998; Meccas and Strauss 1996; Winstanley *et al.* 1999), which are made up of a number of homologous proteins with export functions and are involved in delivering virulence factors directly to host cells. The two type III secretion gene clusters of *B. pseudomallei* were discovered by sequence homology to known type III secretion apparatuses of other pathogens.

However, an efficient computational tool for identification of virulence factors in the *B. pseudomallei* genome is lacking. Karlin proposed five criteria for detecting PAIs in bacterial genomes, which are based on G+C frequency, genome signature profile, codon biases from the complete genome and extreme of amino acid usage in the proteome (Karlin 2001). To capitalize on the sequence information with regards to our understanding of virulence factors and PAIs, we predicted some putative PAIs regions computationally based on the five parameters proposed by Karlin. The prediction results were manually checked with the general features of PAIs (Hacker and Kaper 2000). This will provide us with useful information on the more likely candidate genes involved in the pathogenesis of *B. pseudomallei* and will facilitate the research on its virulence and the development of vaccines against melioidosis.

Further more, this project will also help to evaluate the feasibility of using computational methods to predict putative protein functions and properties from nucleotide and amino acid sequences.

## **4.2 Materials and Methods**

### **4.2.1 Materials**

*B. pseudomallei* genome sequences were downloaded from the ftp server at the Sanger Institute. The ORFs were based on the genome database constructed in Chapter 2. The ribosomal protein, transcription factor, and chaperone sequences were also extracted from the database.

### **4.2.2 General approaches**

A package of Java programs were written to analyze the G+C frequency, genome signature profile of DNA fragments and codon usage and amino acid usage of each ORF, and to compare them with the genome average value, and the average values of ribosomal protein genes, transcription factor genes and chaperone genes. The calculations were based on Karlin's review (Karlin 2001), with several little modifications. These analyses were usually based on a sliding window size of 50 kbp or 20 kbp, and the sliding step size was 1kb.

### **4.2.3 GC frequency comparison**

GC frequency was calculated based on the sliding window  $W$ , using the following formulae

$$f_{GC} = \frac{\text{Number of G} + \text{Number of C}}{\text{Number of bases}}$$

To compute the average chromosome GC frequency, the window  $W$  was set to be the whole chromosome.

#### 4.2.4 Genome signature contrast

Genome signature profile consists of the array of dinucleotide relative abundance values

$$\rho_{xy}^* = \frac{f_{xy}}{f_x * f_y}$$

Where  $f_{xy}$  is the frequency of the dinucleotide XY, and  $f_x$  and  $f_y$  are the frequency of X and Y respectively. As there are four types of nucleotides (A, T, C and G), there are 16 such  $\rho_{xy}^*$  values for any window W and the whole genome.

The genomic signature contrast between two sequences  $f$  (DNA fragment in window W) and  $g$  (genome) is the average absolute dinucleotide relative abundance difference calculated as:

$$\delta^*(f, g) = (1/16) \sum |\rho_{xy}^*(f) - \rho_{xy}^*(g)|$$

where  $\sum$  sums up the genome signature difference  $\rho_{xy}^*(f) - \rho_{xy}^*(g)$  between each sliding window and the entire chromosome for all the 16 types of XY dinucleotides, and outputs the genome signature contrast  $\delta^*(f, g)$  for every sliding window.

#### 4.2.5 Codon usage contrast (codon bias)

If an amino acid Z is encoded by  $n$  types of codons, and the frequency of the usage of each type of codon is  $c_1, c_2, \dots, c_n$ , the codon usage of amino acid Z can be represented by an array of  $n$  numbers Cz ( $c_1, c_2, \dots, c_n$ ), and  $c_1 + c_2 + \dots + c_n = 1$ .

The condon usage contrast between two gene groups F (ORFs in the sliding window W) and G (ORFs in the whole chromosome) is

$$B(F/G) = \sum_a \left[ P_a(F) \left[ \sum_{(x,y,z)=a} |f(x,y,z) - g(x,y,z)| \right] \right]$$

Where  $P_a(F)$  are the average amino acid  $a$  frequencies of the gene group F;  $\left[ \sum_{(x,y,z)=a} |f(x,y,z) - g(x,y,z)| \right]$  is the sum of codon usage of amino acid  $a$  absolute difference between gene group F and G; and  $\sum_a$  is the sum of product of  $P_a(F)$  and  $\left[ \sum_{(x,y,z)=a} |f(x,y,z) - g(x,y,z)| \right]$  for all amino

acids.

#### **4.2.6 Amino acid contrast**

Divergence in amino acid usage (amino acid contrast) between gene group F and G was calculated by

$$A(F/G) = \sum_a |a(F) - a(G)|$$

Where  $a(F/G)$  is the frequency of amino acid  $a$  in gene group  $F/G$ , and  $\sum_a$  sums up the amino acid contrast for all the 20 amino acids.

#### **4.2.7 Putative alien (pA) gene clusters**

The putative alien gene clusters were calculated based on the codon usage contrast between a gene  $g$  and all ribosomal protein (RP), chaperone (CH), and transcriptional factors (TF). A gene is considered as putative alien if the codon bias  $B(g|RP)$ ,  $B(g|CH)$ ,  $B(g|TF)$ , and  $B(g|C)$  all exceed  $M+0.15$ , where  $M$  is the median codon bias of  $B(g|C)$  over all genes. A cluster of at least 5 pA genes is recognized as a pA gene cluster (one gene gap is allowed). The ribosomal protein, chaperone protein, and transcriptional factors were extracted from the *B. pseudomallei* genome database constructed in Chapter 2.

#### **4.2.8 Positive and negative control**

The currently identified two type III secretion gene clusters of *B. pseudomallei* were used as positive control to validate the above analysis. Random generated sequences were used as negative control.

### **4.3 Results**

#### **4.3.1 Implementation of Karlin's 5 criteria**

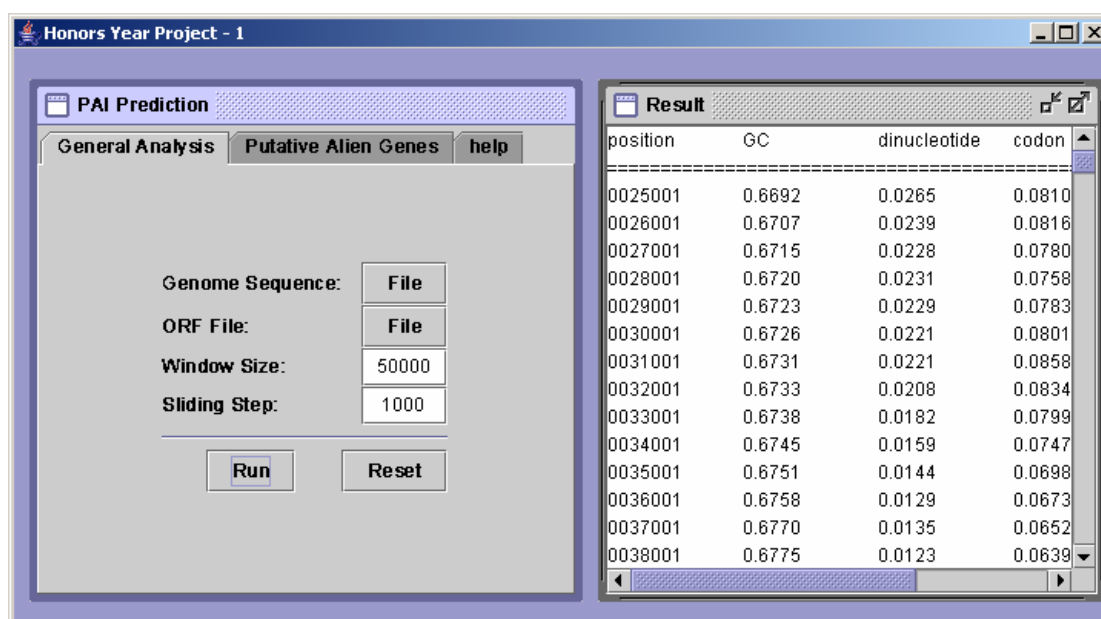
A Java program package with Graphical User Interface (GUI) was written for the PAI and pA gene clusters prediction based on Karlin's 5 criteria. There are 5 files in the source codes of the package. The roles of each of the five files were shown in Table 4.1.

**Table 4.1 Files in the PAI prediction program package**

File	Roles
<b>pai.java</b>	The graphical interface of the program. Take inputs, calculate results using <b>gcodon.java</b> and <b>pA.java</b> , and output results.
<b>gcodon.java</b>	The calculator for GC frequency, genomic signature (dinucleotide frequency), codon usage bias, and amino acid usage bias.
<b>pA.java</b>	The calculator for putative alien genes, based on the codon bias of each ORF with whole proteome, ribosomal proteins, chaperone proteins, and transcription factors.
<b>readFASTA.java</b>	The DNA sequence file (raw sequence or FASTA format) reader, used by both <b>gcodon.java</b> and <b>pA.java</b>
<b>ORFreader.java</b>	The ORF file reader, used by both <b>gcodon.java</b> and <b>pA.java</b>

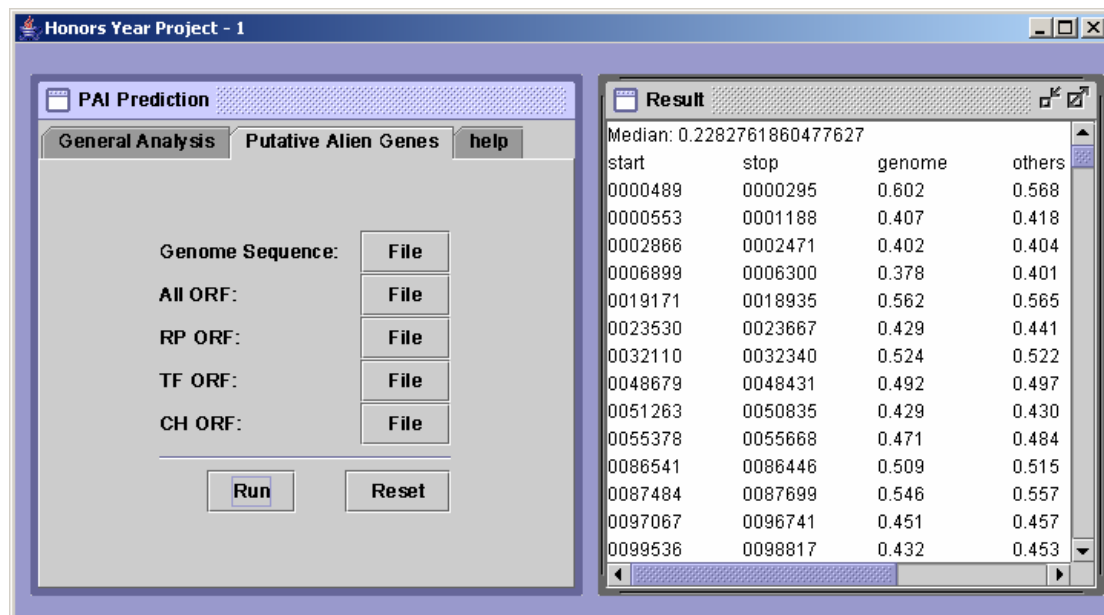
The file **pai.java** is the GUI of the package. The calculations are carried out by **gcodon.java** and **pA.java** in this package, and the calculations are as described in the Section 4.2. The use of this package is simple. For PAI prediction, user need choose genome sequence file and ORF index file, and key in the window size and sliding step for the calculations. Then the program will output the GC frequency, genomic signature, codon usage bias, and amino acid usage bias for each sliding window. Putative PAIs can be identified by peaks of these parameters in the graphs of these four parameters.

**Figure 4.1 Graphical User interface of PAI prediction package for the PAI prediction.**



For putative alien gene clusters prediction, user need choose genome sequence file and ORF files for the whole proteome, ribosomal proteins, chaperone proteins, and transcriptional factors. The program will output indexes of ORFs that have codon usage biases over the four groups of proteins greater than  $M+0.15$ , where  $M$  is the median ORF codon usage bias over the whole proteome.

**Figure 4.2 GUI of the PAI prediction package for the pA gene cluster prediction.**



This package was used to predict of PAIs and pA gene clusters in the *B. pseudomallei* genome. The analysis results were shown in section 4.3.2. The positive and negative controls of the analysis using this package were shown in section 4.3.3.

### 4.3.2 Prediction of PAIs in *B. pseudomallei*

#### 4.3.2.1 PAIs on chromosome 1

The program in section 4.3.1 was used to do pathogenicity islands prediction. The prediction results were shown in Figure 4.3. Five peaks were identified from these four graphs, annotated as A-E. Not all the peaks can be found in all the four graphs. There are only three peaks in the codon usage bias graph, for example. Further analysis of the peaks using

Hacker and Kaper's description are shown in Table 4.2.

**Figure 4.3 PAI prediction on chromosome 1.** Window size of 100kbp and sliding step of 5kbp was used. Putative PAIs were annotated as A-E

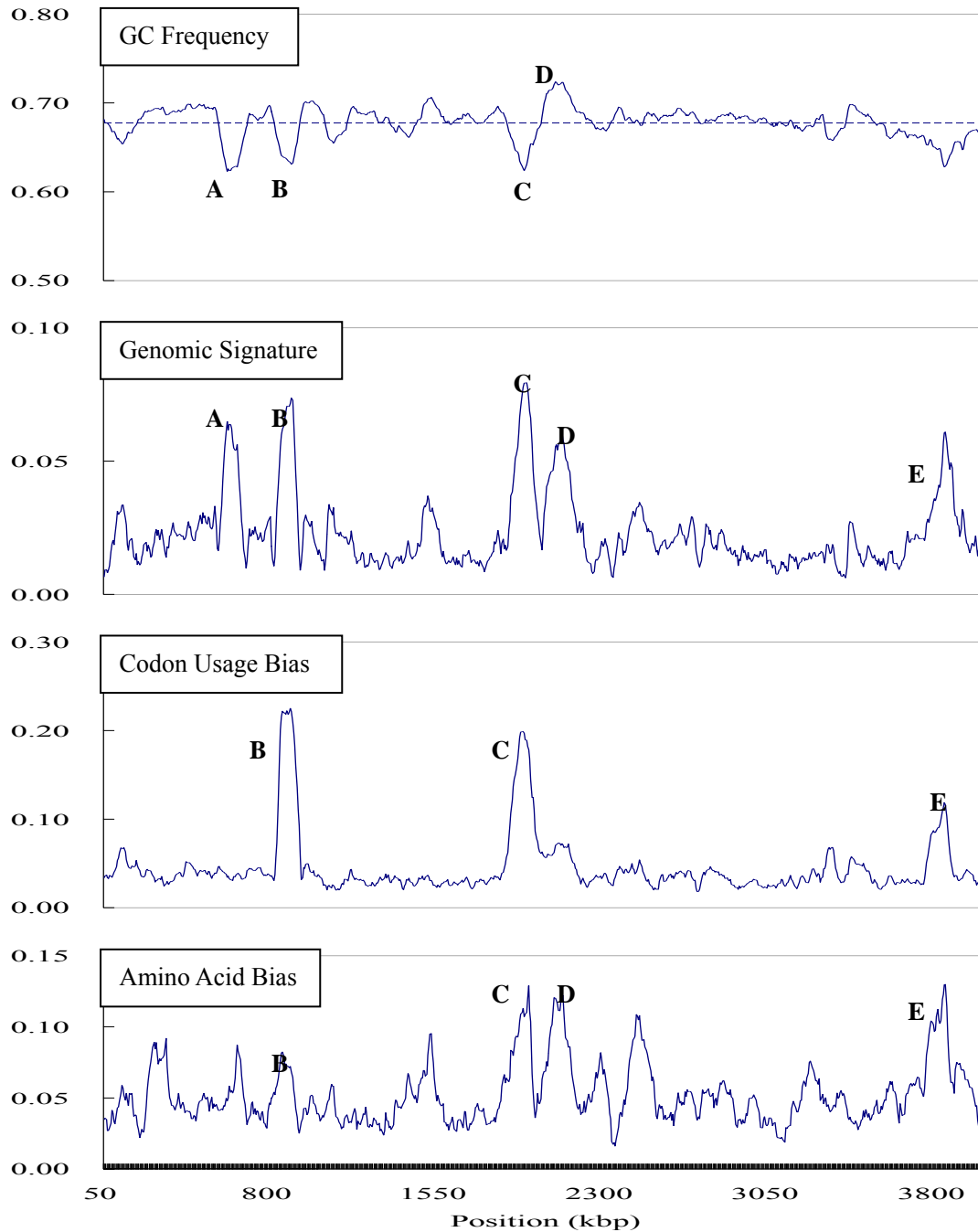


Table 4.2 shows the further analysis results of these peaks on chromosome 1. Peak A was shown obviously in the graph of GC frequency and genome signatures, but not on the graphs of codon usage bias and amino acid bias. The reason is probably that the codon usage bias and amino acid usage bias are based on ORFs, but there are too few predicted ORFs in this

region (11 ORFs over 58kbp). This peak contains putative phage integrase. The functions of most of the genes in this peak are unknown. We called this peak as PAI-1.

**Table 4.2 The PAI candidates on chromosome 1**

Code	PAI	Location (kbp)	Size (kbp)	Gene Accession	Genes	Mobility Genes
A	PAI-1	595-670	75	BP00540-BP00560	Hypothetical proteins	BP00551 Integrase
B	PAI-2	845-910	65	BP00696-BP00727	Prophage proteins Plasmid proteins DNA recombination proteins	BP00702 Integrase
C	PAI-3	1840-2040	200	BP01528-BP01651	Fimbrial proteins Transporters	BP1581 Transposase BP0193 Transposase
D	Not	2080-2110	30	BP01686-BP01701	Cobalamin biosynthesis Magnesium chelatase	
E	Not	3780-3850	70	BP03092-BP03155	Ribosomal proteins	

Peak B is a more typical PAI than peak A. It can be found from all the four graphs. There are many prophage and plasmid proteins in this peak. Putative phage integrase can also be found in this region. We called this peak as PAI-2.

Peak C is another typical PAI. It can be found in all the four graphs, and contains putative transposases. The genes in this region include fimbrial proteins and transporters. We called this peak as PAI-3.

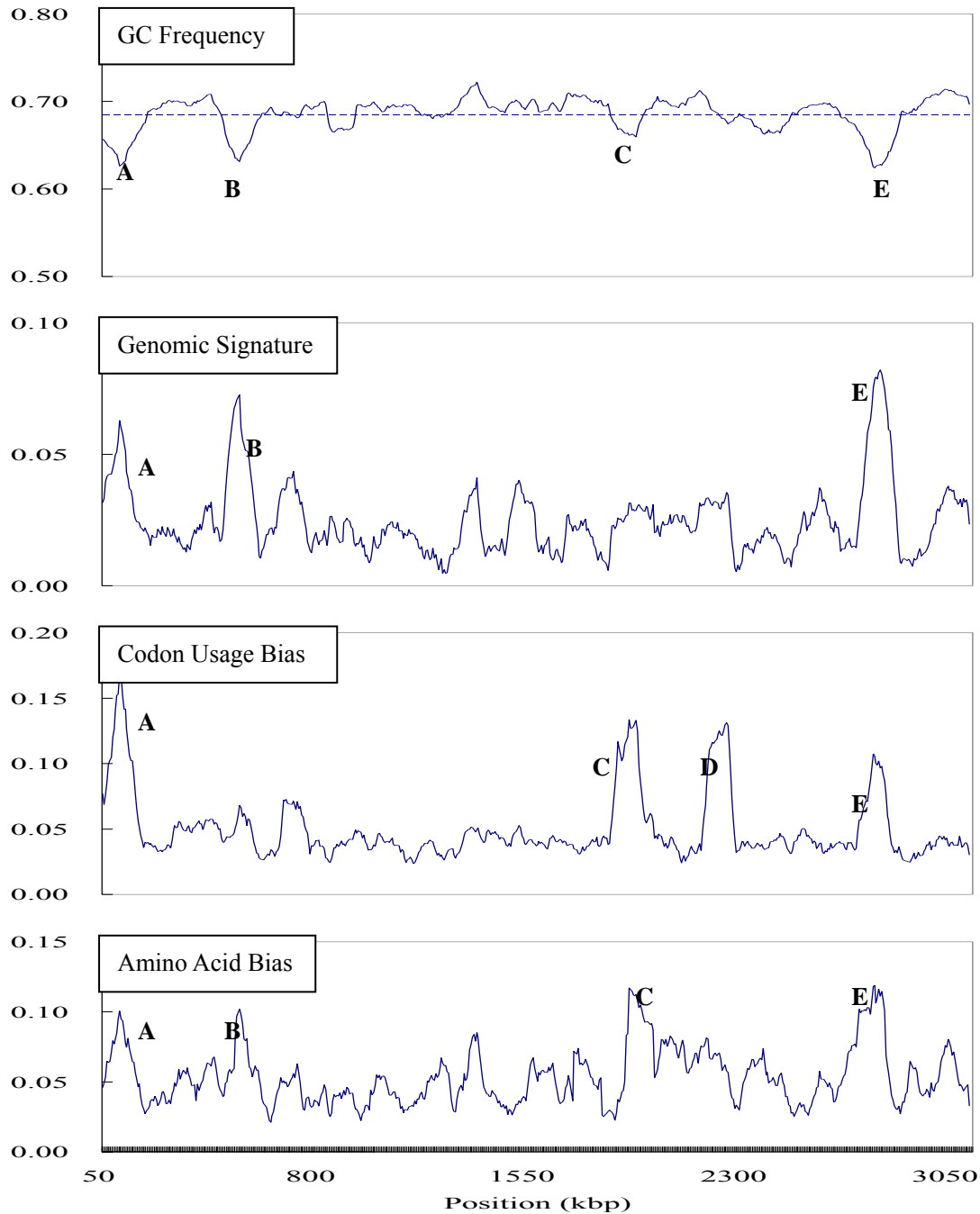
Peak D and E are not PAIs, though they are peaks in some graphs. The size of peak D is too small to be defined as PAI. The dominant genes in peak E are ribosomal proteins, which usually have different codon usage from other genes (Karlín 2001). Therefore, both of them are not PAIs.

#### 4.3.2.2 PAIs on chromosome 2

The prediction results of PAI on chromosome 2 are shown in Figure 4.4. There are 5 peaks can be found from the 4 graphs. Further analysis of each of the peaks using Hacker and

Kaper's description (2000) are shown in Table 4.3.

**Figure 4.4 PAI prediction on chromosome 2.** Window size of 100kbp and sliding step of 5kbp was used. Putative PAIs were annotated as A-E



Peak A can be found in all the four graphs. It contains many outer membrane proteins, mainly fimbrial proteins, which are common PAI proteins. We called this peak as PAI-4.

Peak B can not be found from the codon usage bias graph, and it contains mainly metabolic enzymes. Therefore, it can not be called a PAI, though it is probably have foreign

origin. It may be a genomic island that was horizontal transferred and give *B. pseudomallei* some metabolic advantages, but it is not PAI due to lack of pathogenicity genes.

Peak C and peak D can be found in codon usage bias graph and amino acid usage bias graph. The gene contents of the two peaks are mainly type III secretion systems, which are typical genes in PAI. They are called PAI-5 and PAI-6.

Peak E can be found in all the four graphs. It contains outer membrane proteins and five transposases. These data strongly suggest that peak E is a PAI. We called it as PAI-7.

**Table 4.3 The PAI candidates on chromosome 2**

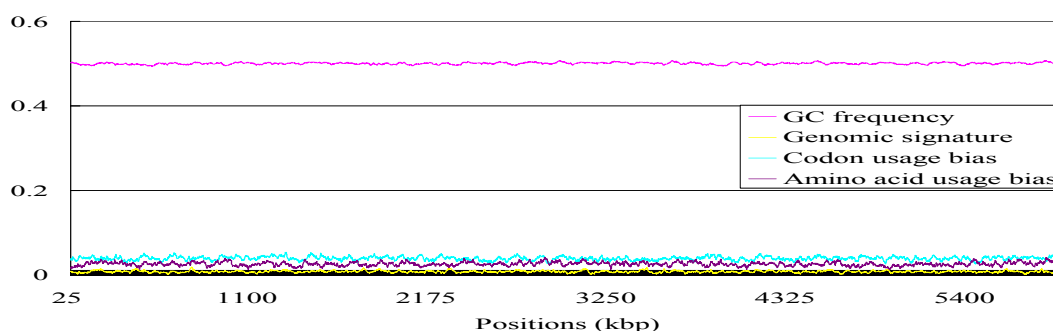
Code	PAI	Location (kbp)	Size (kbp)	Gene Accession	Genes	Mobility Genes
A	PAI-4	80-160	80	BP03395-BP03431	Outer membrane proteins	
B	Not	525-590	65	BP03705-BP03732	Metabolic enzymes	
C	PAI-5	1880-1970	90	BP04704-BP04778	Type III secretion system	
D	PAI-6	2205-2265	60	BP04964-BP04981	Type III secretion system	
E	PAI-7	2760-2825	65	BP05386-BP05401	Outer membrane proteins Transposases	Five Transposases, (active and inactive)

#### 4.3.2.3 Controls of the predictions

Random generated sequences and ORFs were used as negative controls for the analysis.

There are no obvious peaks in all the four graphs, as shown in Figure 4.5.

**Figure 4.5 PAI prediction in random generated sequence and ORFs**



The positive controls of the PAI prediction are the two type III secretion systems. In *B. pseudomallei*, PAIs have been associated with type III secretion systems (Attree and Attree 2001; Hueck 1998; Meccas and Strauss 1996; Winstanley *et al.* 1999), which are made up of a number of homologous proteins with export functions and are involved in delivering virulence factors directly to host cells. The two type III secretion gene clusters of *B. pseudomallei* were discovered by sequence homology to known type III secretion apparatuses of other pathogens. In this analysis, the two type III secretion systems were predicted by the PAI prediction package, as PAI-5 and PAI-6.

### 4.3.3 Putative alien gene clusters

Putative alien (pA) genes of *B. pseudomallei* were predicted using the program package from section 4.3.1, in order to investigate the possible foreign genes that were horizontal transferred into the *B. pseudomallei* genome. The results are shown in Table 4.4.

**Table 4.4 Putative alien (pA) genes in *B. pseudomallei* genome**

	pA genes	% of pA	Known function	Transposon or phage genes	pA clusters (>=5)
Chromosome 1	358	10.8%	123	17	15
Chromosome 2	306	13.0%	112	15	14
Genome	664	11.7%	235	32	29

From the results we can see that there are 11.7% of genes in the *B. pseudomallei* genome were putative alien. This ratio is higher than that of *E. coli*, 7% (Karlin 2001). Most of the pA genes (64.6%) were functionally unknown. This portion is significantly higher than the portion of functionally unknown genes over the whole genome (37.4%). pA gene clusters were also analyzed. The pA gene clusters on chromosome 1 and 2 are shown in Table 4.5 and 4.6 respectively.

**Table 4.5 Putative alien (pA) gene clusters in chromosome 1**

Genes of pA cluster	Number of genes	Function of genes	Transposon or phage genes	PAI association
BP00085-BP00090	6	Hypothetical proteins	BP00088	
BP00125-BP00130	6	Hypothetical proteins	BP00130	
BP00205-BP00210	6	Hypothetical proteins		
BP00549-BP00558	10	Hypothetical proteins	BP00551	PAI-1
BP00701-BP00715	15	Prophage, plasmid, and DNA recombination proteins	BP00703 BP00704 BP00705	PAI-2
BP00995-BP01002	8	Hypothetical proteins		
BP01081-BP01085	6	Hypothetical proteins	BP01082	
BP01314-BP01321	8	Hypothetical proteins		
BP01537-BP01542	6	Hypothetical proteins		PAI-3
BP01574-BP01582	9	Hypothetical proteins	BP01581	PAI-3
BP01593-BP01603	11	Outer membrane proteins	BP01593 BP01603	PAI-3
BP01619-BP01629	11	Hypothetical proteins		
BP02481-BP02485	5	Hypothetical proteins		
BP02713-BP02721	9	Outer membrane proteins		
BP03243-BP03248	6	Hypothetical proteins	BP03243	

Table 4.5 shows that there are 15 pA gene clusters in chromosome 1. Two of them encode outer membrane proteins, and one encodes prophage, plasmid, and DNA recombination proteins. All the other pA gene clusters are functionally unknown. Five of them are located within three PAIs, and eight of them are associated with 11 transposases or phage integrases.

The pA gene clusters of chromosome 2 were shown in Table 4.6. There are 14 pA gene clusters in chromosome 2. Six of them have genes with known functions. Two clusters encode outer membrane proteins, one encodes transposases, and another one encodes type III secretion systems. They are all located within the putative PAIs. Another two clusters encode putative efflux pumps. There is less association with transposon and phage genes for these pA gene clusters than that of chromosome 1 clusters.

**Table 4.6 Putative alien (pA) gene clusters in chromosome 2**

Genes of pA cluster	No. of genes	Function of genes	Transposon or phage genes	PAI association
BP03400-BP003406	7	Outer membrane proteins		PAI-4
BP03424-BP03430	7	Outer membrane proteins		PAI-4
BP03532-BP03536	5	Hypothetical proteins		
BP03706-BP03711	6	Hypothetical proteins		
BP03839-BP03846	8	Efflux pumps		
BP04104-BP04109	6	Hypothetical proteins		
BP04359-BP04370	12	Hypothetical proteins		
BP04675-BP04679	5	Efflux pumps		
BP04720-BP04738	19	Type III secretion system		PAI-5
BP04758-BP04774	7	Hypothetical proteins		PAI-5
BP05257-BP05262	6	Hypothetical proteins		
BP05343-BP05348	4	Hypothetical proteins	BP05348	
BP05385-BP05386	5	Transposases	BP05386, BP05388	PAI-7
BP05396-BP05401	6	Hypothetical proteins	BP05396, BP05461	PAI-7

From these results, we can see that six of the seven predicted PAIs contain one or more pA gene clusters. Therefore, we are more confident for the prediction of PAIs at part 4.3.2.

#### **4.4 Discussion**

Microbial genomes are substantially homogeneous in G+C ratio, codon usage, and genome dinucleotide signature (Karlin 2001), but the composition of pathogenicity islands (PAIs) differs from that of the overall genome. This is due to the PAIs are horizontal transferred into the genome from other genomes with different genomic profile. Using this hypothesis, Karlin proposed 5 criteria for detecting of PAIs and putative alien gene clusters. When a genomic segment deviates sufficiently from the average genome composition in GC frequency, and/or genome signature, and/or codon usage, and/or amino acid frequencies and/or contains a cluster of pA genes, we view the gene segment as a genomic island, or pathogenicity island if there are pathogenic genes (Karlin 2001). If the same segment is identified using all five criteria, we are more confident in the PAI prediction.

In this study, we developed a program package to implement Karlin's 5 criteria, and

predicted the putative PAIs of *B. pseudomallei* genome. The PAI candidates were manually checked according the PAI features described by Hacker and Kaper (2000). Finally, 7 putative PAIs were identified. Two previously known PAIs in *B. pseudomallei*, which are type III secretion systems, are amongst the 7 which we have predicted. Thus, the prediction is reliable. There is an interesting phenomenon, that almost all the PAIs have a G+C ratio smaller than the genome average. This is not due to the high G+C ratio of *B. pseudomallei*, because this phenomenon can also be found in many other low G+C ratio genomes (data not shown). To explain this phenomenon, we propose that low G+C ratio DNA fragment is easier to integrate into the host genome. However, this hypothesis is hard to prove.

Putative alien (pA) gene clusters were also predicted using the codon usage bias for each of the ORFs over four groups of proteins, the whole proteome, the ribosomal proteins, the chaperones, and the transcription factors. 29 pA gene clusters were predicted. 11 of them are located within 6 PAIs, and support the PAI prediction results.

PAIs are usually clusters of pathogenic genes (Hacker and Kaper 2000). For the 7 predicted PAIs, 3 PAIs encode outer membrane proteins, such as fimbrial proteins. These genes have functions for bacteria adhesion to the host, therefore pathogenic. Two predicted PAIs encode type III secretion systems, which have been proven as a major component of bacteria virulent. Another one PAI encodes plasmid, phage, and DNA recombination proteins, and the last one encodes mostly functionally unknown proteins. PAIs are usually present in the pathogenic bacteria, but absent in its nonpathogenic close relevant strains. Therefore, experimental approach to confirm the PAI prediction results would be the screening of non- or less pathogenic strains of *B. pseudomallei* for the existence of the predicted PAIs. If the predicted PAIs are indeed absent in non- or less pathogenic strains of *B. pseudomallei*, then the PAI prediction will be proven.

## CHAPTER 5. DISCUSSION AND CONCLUSION

The *Burkholderia pseudomallei* genome was sequenced by the Sanger Institute on May 2003. The genome comprises a 4.1 Mbps and a 3.2 Mbps replicon, which are called chromosome 1 and chromosome 2, respectively. The *B. pseudomallei* genome structure is interesting, firstly because it is larger than most of other known microbial genomes, and secondly, because it has two large replicons, while most other microbial genomes have a single large replicon. Since the larger replicon in *B. pseudomallei* is of similar size with most other bacteria, such as *E. coli*, we wondered if the smaller replicon might be a megaplasmid. In order to investigate this hypothesis, we annotated the *B. pseudomallei* genome and stored the information in a relational database and applied this database to further analyses of the *B. pseudomallei* genome.

The ORFs were first predicted and each translated ORF was analyzed by homology and conserved domain search. This allowed the ORFs to be classified into many functional categories. Several biochemical pathways in *B. pseudomallei* were also identified. A comparison of genes encoding such biochemical pathways showed several differences between the two replicons.

From the homology search, we found that the smaller replicon is much more diverse than the larger one. 49.4% ORFs of the larger replicon have best homology with genes in *Ralstonia solanacearum*, while only 4.6% have best homology with its second best similar organism, *Pseudomonas aeruginosa*. The smaller replicon is much more diverse in terms of phylogenetic origin, only 19.2% ORFs have best homology with its best similar organism, *R. solanacearum*. This suggests that the smaller replicon may be a plasmid which had evolved by taking genes from very diverse sources.

Then we studied the origins of replication of the two replicons and found that the origin of replication of the smaller replicon is similar to that of a plasmid. This DNA sequence at the origin of chromosome 2 encodes a plasmid replication initiator protein, RepA, and contains 15 direct repeats, or iterons, in an AT-rich region. They are the putative RepA binding sites. Two more repeats of the iterons are located at the promoter region of the RepA gene, and thus may be function as auto-suppress of RepA expression. Partition proteins for active partition of low copy number plasmid, ParA and ParB, are also encoded by this region. All these features are of typical plasmid origin of replication with theta replication mechanism (del Solar *et al.* 1998). This supports the notion that the smaller replicon is a megaplasmid.

Further analysis of the gene contents of the two replicons shows that, the larger replicon encodes a complete set of essential housekeeping genes including genes required for DNA replication, cell division, transcription, and translation. 52 of the 59 tRNAs are located on the larger replicon. All the ribosomal proteins and 3 of the 4 duplicated rRNA operons are located on the larger replicon too. All essential genes required for purine and pyrimidine biosynthesis are also located on the larger replicon. All the amino acid biosynthesis, coenzyme biosynthesis, electron transport and phosphorylation, and most of the energy production and conversion pathways can be found on the larger replicon, while very few on the smaller replicon. These data shows that the larger replicon itself can support *B. pseudomallei* survival and growth, while the smaller replicon is probably a dispensable genetic element.

Therefore, we think that the smaller replicon of *B. pseudomallei* is a dispensable megaplasmid of 3.2 Mbps. It is larger than the largest known megaplasmid, 2.1Mbps, found in *Ralstonia solanacearum* (Salanoubat *et al.* 2002). The megaplasmid may rise from a small plasmid by taking genes from diverse source. It confers on *B. pseudomallei* a survival

advantage allowing it to colonize diverse environments, ranging from soil, water, plants animals and humans.

Then, the question remains as to why *B. pseudomallei* carry a 3.2 Mbp megaplasmid, if the larger replicon carries all the housekeeping genes? We propose that the megaplasmid probably confers a growth advantage, allowing the bacterium to grow in more diverse environments, ranging from soil and water, to plants, animals and humans. There are significantly more transcriptional regulators and signal transduction systems in the smaller replicon than in the larger replicon, thus *B. pseudomallei* can respond faster and more robust to the environmental stimulus. By comparing the portion of transcriptional regulators and signal transduction systems of the two replicons and the whole genome with other bacteria, we can see that the larger replicon alone is similar with other relatively simpler bacteria, such as *E. coli* and *B. subtilis*, while with the smaller replicon, *B. pseudomallei* becomes similar with more complicated bacteria, such as *P. aeruginosa*. Although the larger replicon carries full sets of housekeeping pathway genes, paralogous grouping results showed that genes in the larger replicon are less redundant. In other words, there is little diversity for the genes in the larger replicon, and therefore, *B. pseudomallei* would only grow in less diverse environments if it only carried the larger replicon. The smaller replicon gave the diversity of functions of the proteome. The smaller replicon carries many genes that did not assemble into complete pathways. These genes may replace its counterparts of the larger replicon and alter the pathways in respond to the changing environments. Furthermore, the distribution of secondary metabolites biosynthesis, transport, and catabolism genes is significantly biased in favour of the smaller replicon. From the GO database, 151 genes in the larger replicon and 173 genes in the smaller replicon are classified as virulent factors, and their distribution is significantly biased in the favour of the smaller replicon ( $p = 9.3e-6$ ). These data also support

the hypothesis that the smaller replicon is the source of the *B. pseudomallei* genome complexity. Therefore, we propose that, without the smaller replicon, *B. pseudomallei* would be a complete, but simpler bacterium like *E. coli*, due to the completeness of the larger replicon genes, however, it would only grow in relatively fewer environments, due to the lack of the smaller replicon genes.

However, because the two replicons have evolved together for such a long time so that their GC ratio and codon usage has adapted similar; there must be many DNA recombinations between the two replicons. Therefore, some of the genes in the two replicons have been mixed. One example of this mixing is the ribosomal DNA loci. There are four identical ribosomal DNA loci in the genome, 3 on the larger replicon and 1 on the smaller replicon. These loci are most probably derived from duplication and distributed by recombination and rearrangement. Therefore, some genes in one replicon may be derived from recombination from the other replicon. This gives us problems for the study of differences between the two replicons.

Experimental approaches are necessary to confirm that the smaller replicon is a dispensable megaplasmid. These will be our future work. Small plasmid will be constructed using the origin of replication of the smaller replicon, to investigate the characteristics of the origin of replication of the megaplasmid. Then we will try to knock out the smaller replicon using the constructed plasmid, based on incompatibility of the two plasmids. The characteristics of megaplasmid-deficient strain of *B. pseudomallei* will be studied.

## References

- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.
- Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29.
- Ashdown, L.R. and J.M. Koehler. 1990. Production of hemolysin and other extracellular enzymes by clinical isolates of *Pseudomonas pseudomallei*. *J Clin Microbiol* **28**: 2331-2334.
- Attree, O. and I. Attree. 2001. A second type III secretion system in *Burkholderia pseudomallei*: who is the real culprit? *Microbiology* **147**: 3197-3199.
- Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Ewinger, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L. Sonnhammer. 2002. The Pfam protein families database. *Nucleic Acids Res* **30**: 276-280.
- Besemer, J., A. Lomsadze, and M. Borodovsky. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607-2618.
- Burns, D.L., J.T. Barbieri, B.H. Iglewski, and R. Rappuoli. 2003. *Bacterial Protein Toxins*. ASM Press, Washington, DC.
- Dance, D.A., V. Wuthiekanun, W. Chaowagul, and N.J. White. 1989. The antimicrobial susceptibility of *Pseudomonas pseudomallei*. Emergence of resistance in vitro and during treatment. *J Antimicrob Chemother* **24**: 295-309.
- del Solar, G., R. Giraldo, M.J. Ruiz-Echevarria, M. Espinosa, and R. Diaz-Orejas. 1998. Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* **62**: 434-464.
- Fisher, R.A. 1922. On the interpretation of chi-square from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85**: 87-94.
- Hacker, J. and J.B. Kaper. 2000. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54**: 641-679.
- Hueck, C.J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* **62**: 379-433.
- Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9**: 335-343.
- Leelarasamee, A. and S. Bovornkitti. 1989. Melioidosis: review and update. *Rev Infect Dis* **11**: 413-425.
- Letunic, I., R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, J. Schultz, C.P. Ponting, and P. Bork. 2004. SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32 Database issue**: D142-144.
- Lobry, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**: 660-665.
- Lowe, T.M. and S.R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955-964.
- Marchler-Bauer, A., J.B. Anderson, C. DeWeese-Scott, N.D. Fedorova, L.Y. Geer, S. He, D.I. Hurwitz, J.D. Jackson, A.R. Jacobs, C.J. Lanczycki, C.A. Liebert, C. Liu, T. Madej, G.H. Marchler, R. Mazumder, A.N. Nikolskaya, A.R. Panchenko, B.S. Rao, B.A. Shoemaker, V. Simonyan, J.S. Song, P.A. Thiessen, S. Vasudevan, Y. Wang, R.A. Yamashita, J.J. Yin, and S.H. Bryant. 2003. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res* **31**: 383-387.

- Mecenas, J.J. and E.J. Strauss. 1996. Molecular mechanisms of bacterial virulence: type III secretion and pathogenicity islands. *Emerg Infect Dis* **2**: 270-288.
- Moore, R.A., D. DeShazer, S. Reckseidler, A. Weissman, and D.E. Woods. 1999. Efflux-mediated aminoglycoside and macrolide resistance in *Burkholderia pseudomallei*. *Antimicrob Agents Chemother* **43**: 465-470.
- Ribeiro De Vasconcelos, A.T., o. 1, and o. 2. 2003. The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. *Proc Natl Acad Sci U S A* **100**: 11660-11665.
- Salanoubat, M., S. Genin, F. Artiguenave, J. Gouzy, S. Mangenot, M. Arlat, A. Billault, P. Brottier, J.C. Camus, L. Cattolico, M. Chandler, N. Choisine, C. Claudel-Renard, S. Cunnac, N. Demange, C. Gaspin, M. Lavie, A. Moisan, C. Robert, W. Saurin, T. Schiex, P. Siguier, P. Thebault, M. Whalen, P. Wincker, M. Levy, J. Weissenbach, and C.A. Boucher. 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415**: 497-502.
- Sexton, M.M., L.A. Goebel, A.J. Godfrey, W. Choawagul, N.J. White, and D.E. Woods. 1993. Ribotype analysis of *Pseudomonas pseudomallei* isolates. *J Clin Microbiol* **31**: 238-243.
- Sexton, M.M., A.L. Jones, W. Chaowagul, and D.E. Woods. 1994. Purification and characterization of a protease from *Pseudomonas pseudomallei*. *Can J Microbiol* **40**: 903-910.
- Stein, L. 2001. Genome annotation: from sequence to biology. *Nat Rev Genet* **2**: 493-503.
- Stover, C.K., X.Q. Pham, A.L. Erwin, S.D. Mizoguchi, P. Warrenner, M.J. Hickey, F.S. Brinkman, W.O. Hufnagle, D.J. Kowalik, M. Lagrou, R.L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L.L. Brody, S.N. Coulter, K.R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G.K. Wong, Z. Wu, I.T. Paulsen, J. Reizer, M.H. Saier, R.E. Hancock, S. Lory, and M.V. Olson. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**: 959-964.
- Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278**: 631-637.
- Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**: 22-28.
- Whitmore, A. and C.S. Krishnaswami. 1912. An account of the discovery of a hitherto undescribed infective disease occurring among the population of Rangoon. *India Med Gaz* **47**: 262-267.
- Winstanley, C., B.A. Hales, and C.A. Hart. 1999. Evidence for the presence in *Burkholderia pseudomallei* of a type III secretion system-associated gene cluster. *J Med Microbiol* **48**: 649-656.
- Yabuuchi, E. and M. Arakawa. 1993. *Burkholderia pseudomallei* and melioidosis: be aware in temperate area. *Microbiol Immunol* **37**: 823-836.
- Yabuuchi, E., Y. Kosako, H. Oyaizu, I. Yano, H. Hotta, Y. Hashimoto, T. Ezaki, and M. Arakawa. 1992. Proposal of *Burkholderia* gen. nov. and transfer of seven species of the genus *Pseudomonas* homology group II to the new genus, with the type species *Burkholderia cepacia* (Palleroni and Holmes 1981) comb. nov. *Microbiol Immunol* **36**: 1251-1275.