Patterns of Large-Scale Genomic Variation in Virulent and Avirulent *Burkholderia* Species

Supplementary Information

Catherine Ong[1,**], Ooi Chia Huey[2,**], Dongling Wang[1], Hweeling Chong[2], Ng Kim Chong[2], Fiona Rodrigues[2], May Ann Lee[1,*], and Patrick Tan[1,2,*]

[1]Defence Medical Research Institute / [2]National Cancer Centre, 11 Hospital Drive, Singapore 169610, Republic of Singapore

** C.O. and O.C.H. contributed equally to this report

* Address correspondence to P.T. (cmrtan@nccs.com.sg) or M.A.L. (mayann@dsta.gov.sg)

Supplementary Data S1: Prediction of *B. pseudomallei* Genes (ORFs) and Microarray Fabrication (Expanded)

The ORF prediction program GLIMMER (v2.0) was used to identify potential ORFs in partially assembled sequence data of *B. pseudomallei* strain K96243 (downloaded from the Sanger Centre in March 2001). Using a minimal gene length of 400 bp, 8776 putative ORFs were initially identified. A preliminary analysis revealed that many of these ORFs were overlapping and redundant, and thus the individual ORFs were consequently grouped into distinct homology 'clusters' by reiteratively BLASTing each ORF against all others. 6895 unique ORF 'clusters' were defined, where each 'cluster' contained 1-3 members of similar directionality and sequence. For each cluster, a representative sequence ("RS") (usually the longest) was chosen, and Primer 3 software was used to design oligonucleotide primer pairs to amplify 300bp-1000bp fragments for each RS. To minimize the possibility of cross-hybridization to related genomic sequences, each amplified fragment was cross-checked for uniqueness for BLASTing its sequence against the entire ORF dataset. Potentially cross-hybridizing sequences were rejected and new sequences were selected in their place.

To compare the results of ORF prediction using partially assembled (PA) and fully assembled (FA) sequences, a new round of ORF prediction was performed using a larger training set (40 sequences in FA vs. 30 in PA) and the fully assembled *B. pseudomallei* genome sequence, resulting in the identification of 8412 putative ORFs. To compare the results from the two prediction runs (PA vs. FA) the 6895 RS's from the PA prediction were tiled against the completed genome to determine the specific chromosomal localization of each sequence. The FA ORF dataset was then queried to determine if it contained an entity at a similar chromosomal location. Approximately 80.5% of the PA set (5553 out of 6895) was mappable to the new FA set (within a tolerance of +/– 100bp). Of the remaining 1322 MC sequences, an additional 756 (11%) could be matched to a member in the FA dataset of at least 50 % sequence similarity. Finally, a reciprocal analysis was performed where the FA ORFs were BLASTed against the 6895 PA ORFs. Using this analysis, only 110 ORFs from the FA ORF dataset failed to be convincingly assigned to a partner entity in the PA ORF dataset (P.T., data not shown).

Microarray Fabrication

Genomic DNA from *B. pseudomallei* strain K96243 was used as a template to PCR amplify array probes, and amplification success was visually assessed using 1.2% agarose gel electrophoresis where each reaction was scored for the presence of a band of the expected size. After PCR purification and resuspension, each array probe was printed in duplicate on commercially available microarray slides (Full Moon Biosystems, USA) using an SDDC-2 Microarrayer (Virtek, USA). As a negative control, DNA primers for the *hilA* gene, encoding the pathogenicity island specific for *Salmonella paratyphi* (gift from Jimmy Loh and Lim Bock, DMRI) were also included on the array.

<u>Supplementary Data S2: Distribution of Array Probes Across the *B. Pseudomallei* Genome</u>

Figure S2a) Frequency of Array Probe Density. The *B. pseudomallei* genome sequence was divided into a series of sequential 10 kb fragments, and the number of array probes/10 kb fragment was plotted as a frequency graph.

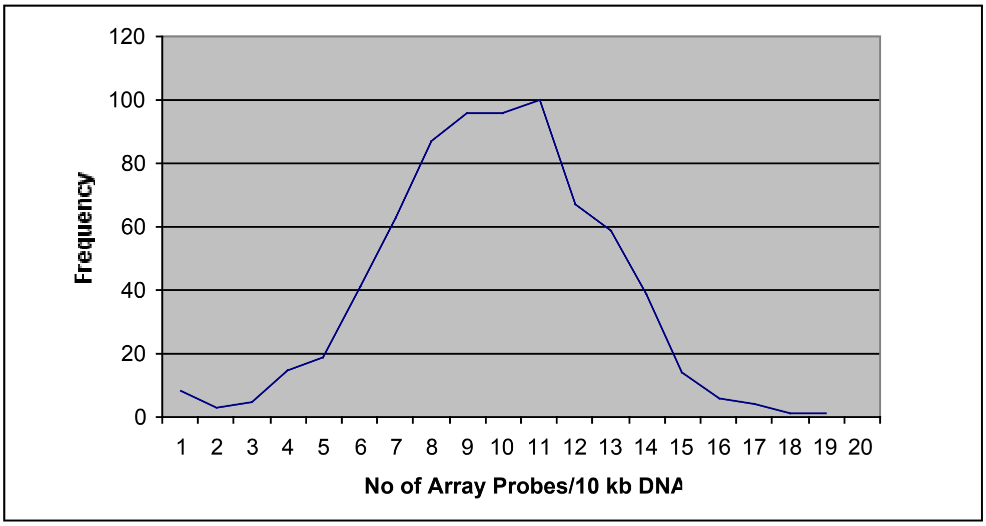
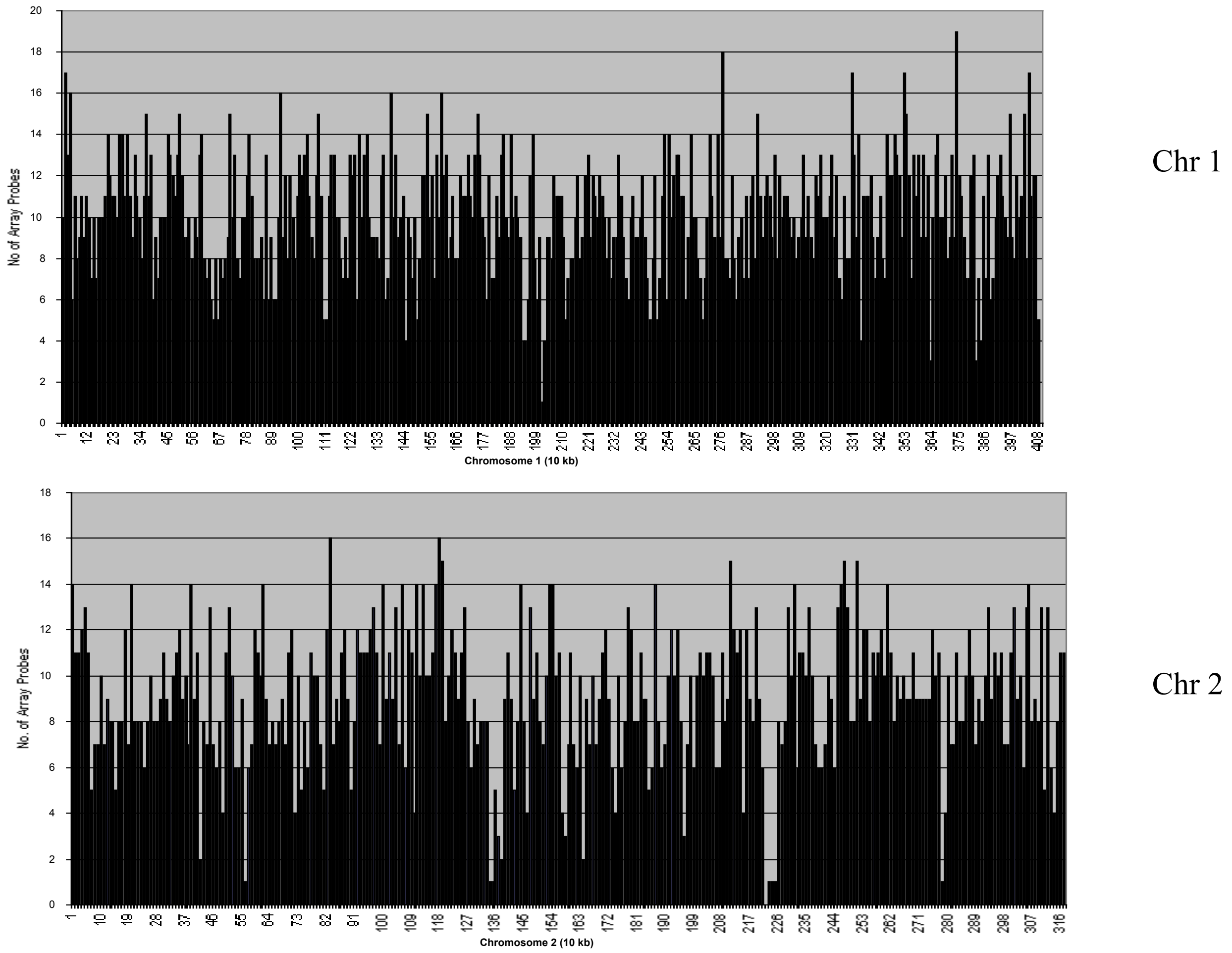


Figure S2b) Array Probe Coverage Along Chromosome 1 and 2. Individual array probes were 'tiled' against the *B. pseudomallei* chromosome 1 (top) and chromosome 2 (bottom) sequences.



Supplementary Data S3: Comparison Between Successful and Failed PCR Reactions

Figure S3a) Frequency of Failed PCR Amplification Reactions Categorized by Predicted PCR Product (pdt) Size

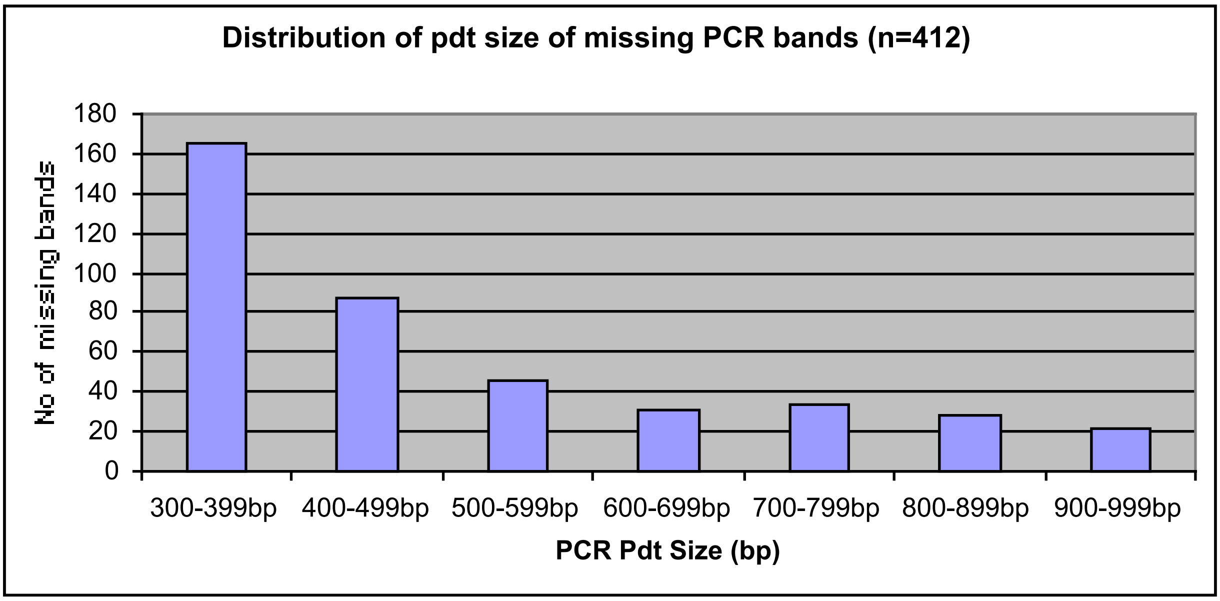
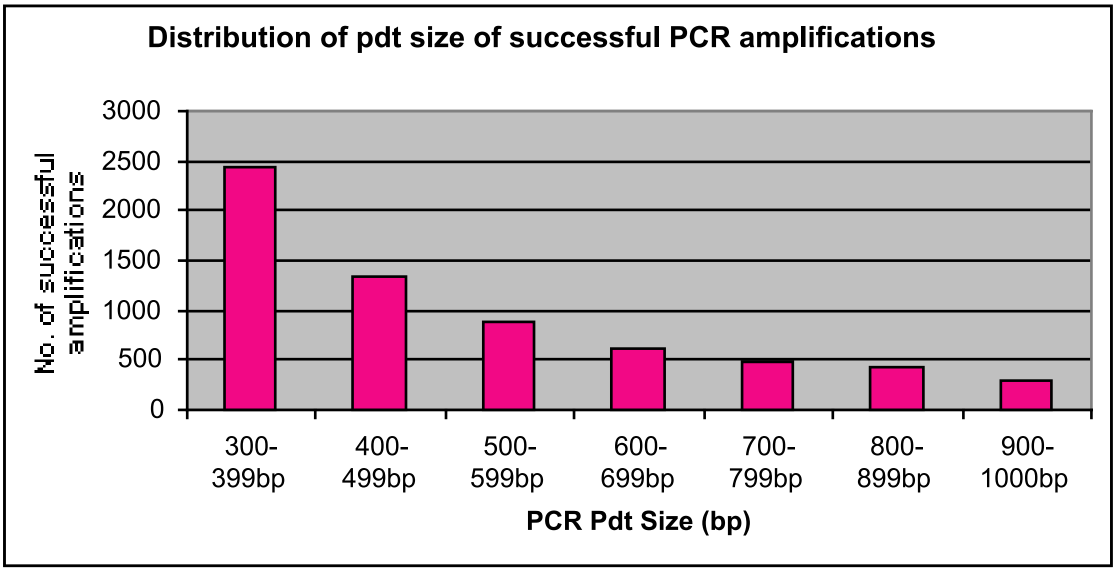


Figure S3b) Frequency of Successful PCR Amplification Reactions Categorized by Predicted PCR Product (pdt) Size



Supplementary Data S4: Comparison of ORF-wise standard deviation among the 3 *Burkholderia* species

Basic stats	***B. pseudomallei***	***B. mallei***	***B. thailandensis***
average	0.156	0.102	0.149
median	0.118	0.069	0.117
mod range	0.0875~0.1	0.0125~0.025	0.0875~0.1

ORF(row)-wise standard deviations were calculated for each ORF, *for each individual species*, then summarised in terms of statistics (average, median and mod). Based on the average and median, the ORF-wise standard deviation can be shown to be generally higher for *B. pseudomallei* compared to the other 2 species, indicating that the fluorescence ratio cut-off threshold of 0.6 used in the manuscript can also be applied to *B. mallei* and *B. thailandensis*. Using 0.6 as a threshold value, in fact, is a stricter form of filtering out false positives for amplification and deletion in *B. thailandensis* and *B. mallei*.

Supplementary Data S5: Verification of Amplified ORFs by Quantitative Real-time PCR

To address the issue of 'amplified' ORFs, quantitative real-time PCR experiments were performed to verify the results obtained from the microarray experiments.

Microarray data associated with two 'amplified' ORFs are shown in Table S5. Microarray experiments revealed ORF 3026602 to be amplified in *B. mallei* at 99.99% but absent in *B. thailandensis* at the confidence threshold; while ORF 3041002 was amplified in both *B. mallei* and *B. thailandensis*, at 99.99% and 99.95% confidence threshold respectively (Table S5).

NCCID	**Annotation**	**Accession No.**	**Mean in *B. pm***	**Mean in *B. m***	**Mean in *B. thai.***	**Remarks**
3026602	TnpB [*Burkholderia mallei*]	gb|AAN64700.1|	0.24809	4.109215	-3.8448	Amplified in *B. mallei* at 99.99%; absent in *B. thailandensis* at 99.99%
3041002	Transposase [*Burkholderia mallei*]	gb|AAG02082.1|	-0.11280	1.79559	1.53525	Amplified in *B. mallei* at 99.99%; amplified in *B. thailandensis* at 99.95%

Table S5a : 'Amplified' ORFs used in quantitative real-time PCR verification.

Quantitative real-time PCR was carried out in an I-Cycler (BioRad) with the same PCR conditions used for the PCR amplification of array probes except for the additional of fluorescent SYBR® Green dye. Figures S4a and S4b show the results obtained for ORF 3026602 and ORF 3041002 respectively.

Fig S5a. Real time PCR amplification with ORF NCCID 3026602.

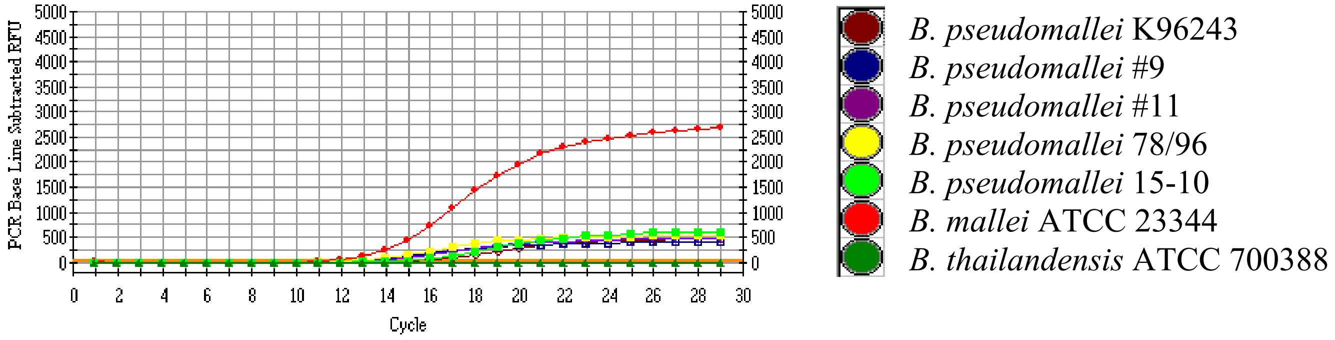


Fig S5b. Real time PCR amplification with ORF NCCID 3041002.

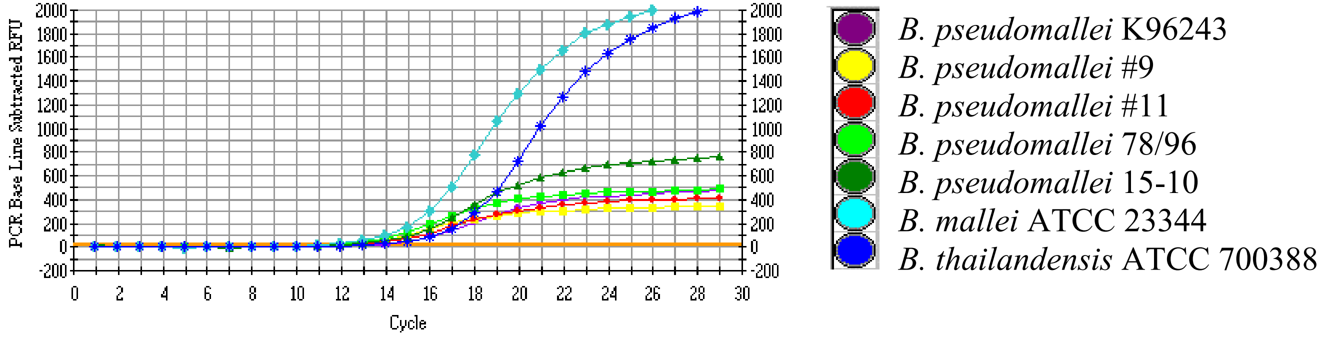


Table S5b : Cross-over values of *B. pseudomallei*, *B. mallei* and *B. thailandensis*.

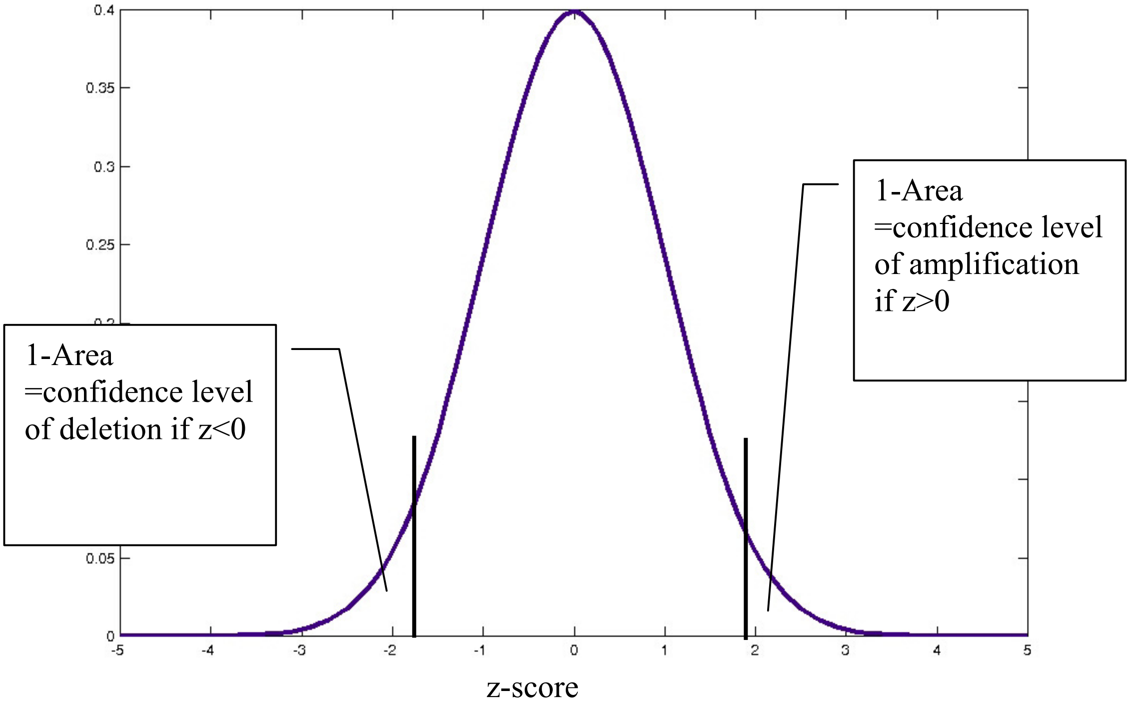
NCCID	**Cross-over values (cycles)**						
	B. pseudomallei **K96243 (ref)**	***B. pseudomallei*** **#9**	***B. pseudomallei*** **#11**	***B. pseudomallei*** **78/96**	***B. pseudomallei*** **15-10**	***B. mallei*** **ATCC 23344**	***B. thailandensis*** **ATCC 700388**
3026602	14.6	11.6	12.4	11.5	14.1	10.9	0.0
3041002	13.7	12.3	12.8	12.0	12.4	11.5	13.2

Real-time PCR of the ‘amplified’ ORFs was performed at least twice to verify the results. The results obtained from both microarray experiment and quantitative real-time PCR show that these ORFs are present in multiple copies in the *B. mallei* genome.

Supplementary Data S6: Intensity-Dependent z-scores Analysis

We have also analysed the microarray data set using the intensity-dependent z-scores methodology. A z-score was calculated for each hybridization value by subtracting the row-wise (ORF-wise) average, and dividing by the row-wise standard deviation. The area under the Gaussian distribution curve between the smallest (largest) Z-value to each negative (positive) z-score was calculated, as shown in Figure S5. For this analysis, we have focused on deletions.

Figure S6: Calculation of confidence level with z-scores analysis



1-Area
=confidence level
of deletion if z<0

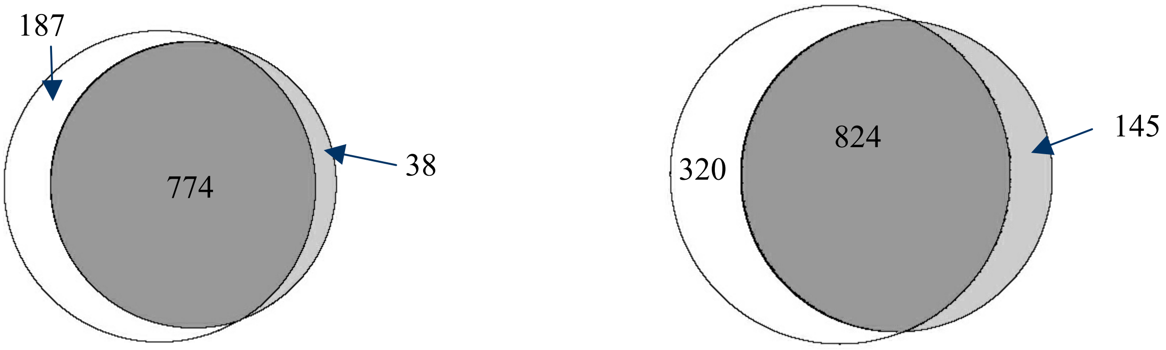
1-Area
=confidence level
of amplification
if z>0

Supplementary Data S7: Comparison of t-test and z-scores Analysis

Table S7: Analysis of Flanking Probe Deletions at >95% and >99.99% confidence levels for B. mallei and B. thailandensis, compared to B. pseudomallei, using z-scores as measure of confidence level (cf Table 2 in Main text)

Species	Coverage	No. of Probes	Probability (%) of Flanking Deletions		
	(% Confidence)		Two Sides	One Side	Isolated
B. mallei	Whole Genome (>95%)	1105	74.84	18.28	6.88
	Whole Genome (>99.99%)	961	79.81	17.27	2.91
	Chr 1 (>95%)	356	65.17	23.60	11.24
	Chr 1 (>99.99%)	302	70.53	24.17	5.30
	Chr 2 (>95%)	749	79.44	15.75	4.81
	Chr 2 (>99.99%)	659	84.07	14.11	1.82
B. thailandensis	Whole Genome (>95%)	1563	41.59	34.17	24.25
	Whole Genome (>99.99%)	1144	51.22	33.04	15.73
	Chr 1 (>95%)	654	36.09	33.03	30.89
	Chr 1 (>99.99%)	461	46.20	32.10	21.69
	Chr 2 (>95%)	909	45.54	34.98	19.47
	Chr 2 (>99.99%)	683	54.61	33.67	11.71

Figure S7: Venn Diagrams of ORFs found to be deleted at 95% and above confidence level using the Student's t-test (lightly shaded areas), and overlap (darker shade) with ORFs found to be deleted at 99.99% and above confidence level using the z-scores method (unshaded areas).



B. mallei

B. thailandensis

Supplementary Data S8a: Reiterative Analysis of Flanking Probe Deletions (Detailed Analysis Using t-test as a Measure of Confidence)

B. mallei (Overall)

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	59.01	58.43	58.47	59.24	1.45	8.25
one sided	33.72	33.95	33.22	32.51	3.65	7.05
none	7.27	7.62	8.31	8.25	94.90	84.69
No. of deletions (accumulated)	344	433	602	812		

***B. thailandensis* (Overall)**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	49.01	45.90	42.43	41.49	1.50	7.12
one sided	33.22	34.43	35.28	35.09	10.40	13.87
none	17.76	19.67	22.29	23.43	88.10	79.01
No. of deletions (accumulated)	304	488	839	969		

***B. mallei* Against *B. pseudomallei* Chr 1**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	45.28	45.38	46.77	48.78	0.84	3.74
one sided	42.45	43.08	41.40	37.40	2.51	4.62
none	12.26	11.54	11.83	13.82	96.65	91.64
No. of deletions (accumulated)	106	130	186	246		

***B. mallei* Against *B. pseudomallei* Chr 2**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	65.13	64.03	63.70	63.78	2.48	14.75
one sided	29.83	30.03	29.57	30.39	5.57	10.54
none	5.04	5.94	6.73	5.83	91.95	74.71
No. of deletions (accumulated)	238	303	416	566		

***B. thailandensis* Against *B. pseudomallei* Chr 1**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	46.79	45.56	42.37	41.84	0.84	4.67
one sided	32.11	32.78	30.53	30.53	7.16	9.35
none	21.10	21.67	27.10	27.63	92.00	85.98
No. of deletions (accumulated)	109	180	321	380		

***B. thailandensis* Against *B. pseudomallei* Chr 2**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	50.26	46.10	42.47	41.26	2.59	10.64
one sided	33.85	35.39	38.22	38.03	15.68	20.33
none	15.90	18.51	19.31	20.71	81.73	69.02
No. of deletions (accumulated)	195	308	518	589		

There is no statistically significant difference in the flanking probabilities across the different confidence intervals ($p>0.05$ for all pairwise comparisons). Flanking probabilities from the last range (>95%) were taken as representative of the 4 ranges in the main text.

Supplementary Data S8b: Reiterative Analysis of Flanking Probe Deletions (Detailed Analysis using z-scores as measure of confidence)

***B. mallei* (Overall)**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	79.81	78.74	76.96	74.84	0.95	12.80
one sided	17.27	17.56	18.07	18.28	4.22	6.47
none	2.91	3.69	4.97	6.88	94.83	80.73
No. of deletions (accumulated)	961	1002	1046	1105		

***B. thailandensis* (Overall)**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	51.22	48.74	45.02	41.59	2.53	11.39
one sided	33.04	33.82	34.48	34.17	19.17	22.57
none	15.73	17.44	20.50	24.25	78.30	66.04
No. of deletions (accumulated)	1144	1233	1395	1563		

***B. mallei* Against *B. pseudomallei* Chr 1**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	70.53	69.62	67.27	65.17	0.49	6.15
one sided	24.17	24.05	24.62	23.60	3.45	5.21
none	5.30	6.33	8.11	11.24	96.06	88.64
No. of deletions (accumulated)	302	316	333	356		

***B. mallei* Against *B. pseudomallei* Chr 2**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	84.07	82.94	81.49	79.44	1.78	22.36
one sided	14.11	14.58	15.01	15.75	5.58	8.28
none	1.82	2.48	3.51	4.81	92.64	69.37
No. of deletions (accumulated)	659	686	713	749		

***B. thailandensis* Against *B. pseudomallei* Chr 1**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	46.20	43.54	40.25	36.09	1.79	7.30
one sided	32.10	33.00	33.04	33.03	14.60	17.56
none	21.69	23.46	26.71	30.89	83.62	75.14
No. of deletions (accumulated)	461	503	569	654		

***B. thailandensis* Against *B. pseudomallei* Chr 2**

deletion at	**>99.99%**	**>99.9%**	**>99%**	**>95% (ALL DELETION)**	**≤95% (NON DELETION)**	**ALL ORFs**
both sides	54.61	52.33	48.31	45.54	3.86	17.26
one sided	33.67	34.38	35.47	34.98	27.25	29.74
none	11.71	13.29	16.22	19.47	68.89	53.01
No. of deletions (accumulated)	683	730	826	909		

There is no statistically significant difference in the flanking probabilities across the different confidence intervals ($p>0.05$ for all pair-wise comparisons). Flanking probabilities from the last range (>95%) were taken as representative of the 4 ranges in the main text.

Supplementary Data S9: Distribution of Deleted ORF Cluster Sizes (using t-test or z-scores, cut-off at ≥95% level)

The Y-coordinates of the plot represent the number of clusters whose sizes are equal to or larger than the values of corresponding X-coordinates (ORF density is taken as a measure of genomic cluster size). Deleted clusters in *B. mallei* are larger than clusters in *B. thailandensis*. Figures 9a represents ORFs identified as deleted by t-test, and Figure 9b by z-scores.

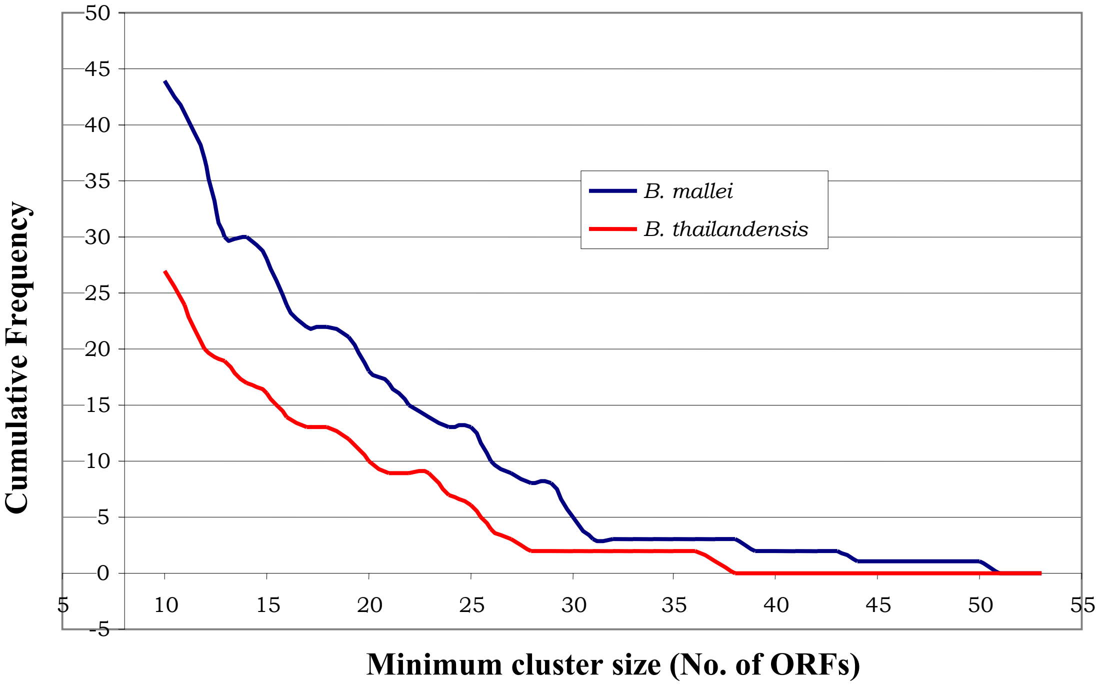


Figure S9a: Frequency Graph of Deleted Clusters by Cluster Size (t-test)

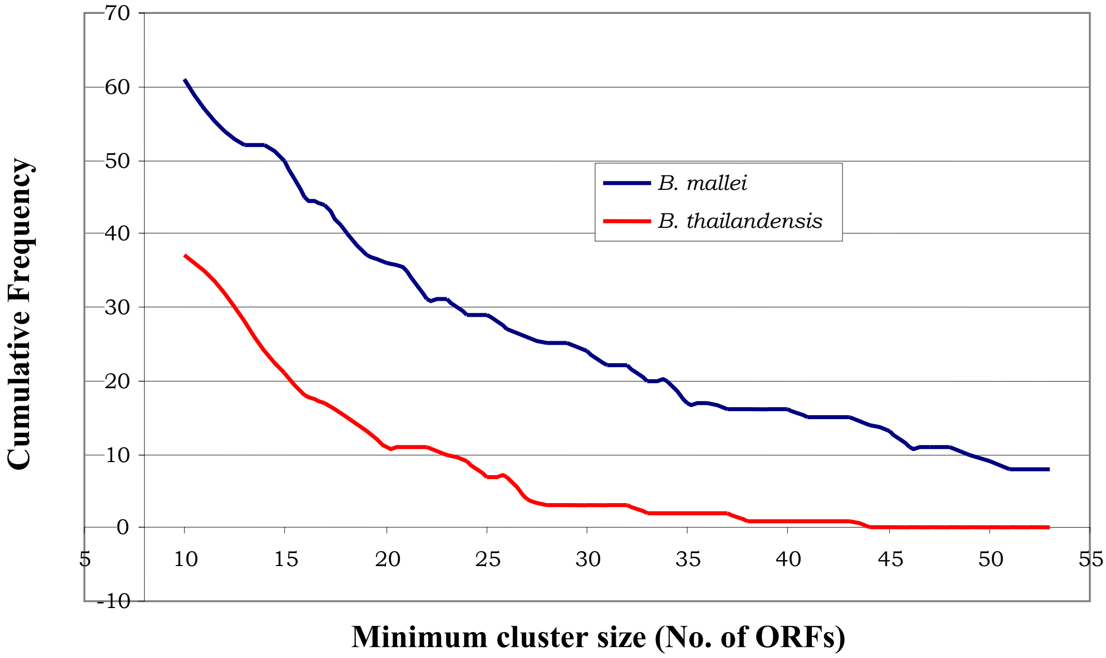


Figure S9b: Frequency Graph of Deleted Clusters by Cluster Size (z-scores)

Supplementary Data S10a: Statistical Analysis of Flanking ORF Deletions in *B. mallei* and *B. thailandensis* (using t-test as measure of confidence)

A chi-square analysis was performed to compare the patterns of ORF deletion in *B. mallei* and *B. thailandensis*. (Bm = *B. mallei*, Bt = *B. thailandensis*, 2s = Deleted ORFs flanking both sides, 1s = Deleted ORFs flanking one side, none = no flanking deleted ORF). The 95% confidence level was used as a cut-off criterion to define a 'deleted' ORF.

Observed

>95%	Bm	Bt	
2s	481	402	883
1s	264	340	604
none	67	227	294
tot	812	969	1781

Expected

2s	402.58	480.42	883
1s	275.38	328.62	604
none	134.04	159.96	294
tot	812	969	1781

Chi-square	
Pvalue	2.15E-20

An Odds-Ratio Calculation was performed to determine the likelihood that a deleted ORF in *B. mallei* would be flanked on both sides by other deleted ORFs (i.e. would belong to a deleted ORF cluster), as compared to *B. thailandensis*.

	ODDS RATIO	
	Bm	Bt
2s	481	402
not 2s	331	567
P(2s)	0.59	0.41
P(not 2s)	0.41	0.59
P(2s)/P(not 2s)	1.45	0.71

An Odds ratio metric of P(2s)/P(not 2s) was obtained for both species

A deleted Bm ORF is **2.05** times more likely to have 2s flanking (Bm/Bt)

Supplementary Data S10b: Statistical Analysis of Flanking ORF Deletions in *B. mallei* and *B. thailandensis* (using z-scores as measure of confidence)

A chi-square analysis was performed to compare the patterns of ORF deletion in *B. mallei* and *B. thailandensis*. (Bm = *B. mallei*, Bt = *B. thailandensis*, 2s = Deleted ORFs flanking both sides, 1s = Deleted ORFs flanking one side, none = no flanking deleted ORF). The 99.99% confidence level was used as a cut-off criterion to define a 'deleted' ORF.

Observed

>99.99%	Bm	Bt	
2s	767	586	1353
1s	166	378	544
none	28	180	208
tot	961	1144	2105

Expected

2s	617.69	735.31	1353
1s	248.35	295.65	544
none	94.96	113.04	208
tot	961	1144	2105

Chi-square
Pvalue 6.35E-45

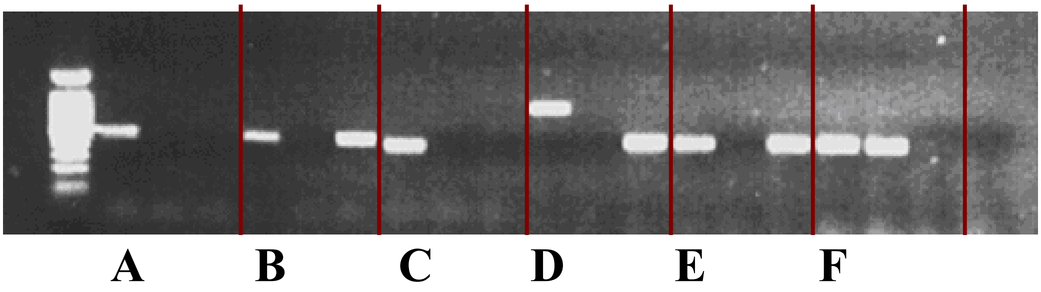
An Odds-Ratio Calculation was performed to determine the likelihood that a deleted ORF in *B. mallei* would be flanked on both sides by other deleted ORFs (i.e. would belong to a deleted ORF cluster), as compared to *B. thailandensis*.

	ODDS RATIO	
	Bm	Bt
2s	767	586
not 2s	194	558
P(2s)	0.80	0.51
P(not 2s)	0.20	0.49
P(2s)/P(not 2s)	3.95	1.05

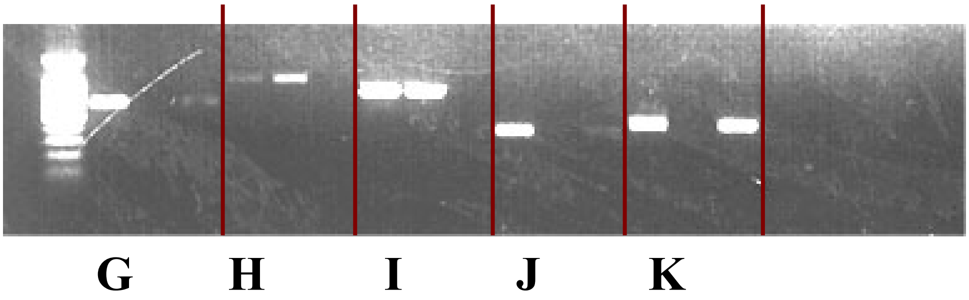
An Odds ratio metric of P(2s)/P(not 2s) was obtained for both species
A deleted Bm ORF is **3.76** times more likely to have 2s flanking (Bm/Bt)

Supplementary Data S11: Experimental Validation of Other Deleted ORFs

Figure S11: PCR amplification reactions were performed using ORF Primer Pairs on three different species of *Burkholderia*. Results are presented in the following arrangement : First lane- *Burkholderia pseudomallei*, Second Lane - *Burkholderia mallei*, Third lane : *Burkholderia thailandensis*



Lanes	ORF Name	Annotation	Deleted in
Set A	ORF 3361902	ABC Transporter Subunit	*B. mallei*
Set B	ORF 2812502	Fe-pyochelin Receptor Precursor	*B. mallei*
Set C	ORF 3174502	Nitrate reductase (delta subunit)	*B. mallei*
Set D	ORF 2847802	Nitrate reductase (beta subunit)	*B. mallei*
Set E	ORF 2847902	Nitrate reductase (alpha subunit)	*B. mallei*
Set F	ORF 3365602	Nitric Oxide Reductase	*B. thailandensis*



Lanes	ORF Name	Annotation	Deleted in
Set G	ORF 3414002	beta-lactamase	*B. mallei*
Set H	ORF 2874802	WbiH (Type II OPS)	*B. thailandensis*
Set I	ORF 3234402	AraC-related Trans Factor	*B. thailandensis*
Set J	ORF 3176002	Quorum Sensing Trans Factor	*B. mallei*
Set K	ORF 3418002	AHL Synthase	*B. mallei*

Supplementary Data S12: Deleted ORFs in *B. mallei* and *B. thailandensis*

NCCID	Chr	Start bp	Stop bp	Annotation	Accession Number	Similarity	Comments
				METABOLISM			
				Deleted in *B. mallei*			
3218202	1	369368	370294	PROBABLE DIHYDRODIPICOLINATE SYNTHASE-RELATED TRANSMEMBRANE PROTEIN	ref|NP_518786.1	200 3e-50	Deleted in *B. thailandensis*
3466202	1	640048	641115	dienelactone hydrolase family protein [Pseudomonas putida KT2440]	ref|NP_743344.1|	149 5e-35	
3452802	1	841255	844113	penicillin acylase II [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_636766.1|	439 e-121	
3054302	1	1450878	1451591	CYC4_PSEAE CYTOCHROME C4 PRECURSOR	sp|P00106	96 4e-19	Operon (Cytochrome)
3054402	1	1453089	1453658	cytochrome-c oxidase (EC 1.9.3.1) chain II [imported] - Bacillus stearothermophilus	pir||T43834	116 2e-25	Operon (Cytochrome)
2818002	1	1453648	1455270	cytochrome c oxidase subunit 1 [Burkholderia pseudomallei]	gb|AAF13732.1|	986 0.0	Operon (Cytochrome)
2785002	1	1999100	2005876	peptide synthetase XpsB [Xenorhabdus bovienii]	gb|AAL57600.1	891 0.0	
2790002	1	2006602	2018016	NosC [Nostoc sp. GSV224] (peptide synthetase)	gb|AAF17280.1	948 0.0	
3394402	1	2304793	2306010	probable lipopolysaccharide modification acyltransferase [Mesorhizobium loti]	ref|NP_102483.1	218 2e-55	
3043102	1	2508356	2509234	DNA adenine methylase [Bacteriophage phiE125]	ref|NP_536383.1|	118 2e-44	
3170302	1	3309224	3310654	2-dehydro-3-deoxyphosphooctonate aldolase [Pseudomonas putida KT2440]	ref|NP_743962.1|	397 e-109	
3274902	2	575780	578404	polysaccharide biosynthesis protein [Shewanella oneidensis MR-1]	ref|NP_718749.1	171 6e-41	
2852502	2	652895	654517	2-aminobenzoate-CoA ligase [Azoarcus evansii]	gb|AAL02069.1|	551 e-155	
3361902	2	810019	811008	probable ATP-binding component of ABC transporter [Pseudomonas aeruginosa PA01]	ref|NP_252912.1|	233 4e-60	Iron Metabolism
2812502	2	811073	813646	Fe(III)-pyochelin receptor precursor [Pseudomonas aeruginosa PA01]	ref|NP_252911.1|	884 0.0	Iron Metabolism
2789502	2	843109	844809	PROBABLE ACETYL-COENZYME A SYNTHETASE (ACETATE--COA LIGASE) PROTEIN [Ralstonia solanacearum]	ref|NP_522212.1|	829 0.0	
3357302	2	1483082	1484089	Chain A, 4-Amino-5-Hydroxymethyl-2-Methylpyrimidine Phosphate Kinase From Salmonella Typhimurium	pdb|1JXI|A	236 4e-61	
3014102	2	1488211	1489524	epsilon-trimethyllysine hydroxylase [Rattus norvegicus]	ref|NP_596878.1	89 1e-16	Deleted *in B. thailandensis*
3319502	2	1509034	1510014	TRP-2 [Burkholderia glumae]	dbj|BAB88915.1|	451 e-126	Riboflavin
2910202	2	1510127	1511311	riboflavin-specific deaminase [Xylella fastidiosa 9a5c]	ref|NP_298240.1|	172 8e-42	Riboflavin
3172302	2	1521595	1522422	hydroxyethylthiazole kinase [Leptospira interrogans serovar lai str. 56601]	ref|NP_713545.1|	124 1e-27	
2847602	2	1532876	1535935	hydrogenase-4 component B [Burkholderia pseudomallei]	gb|AAK06855.1|	1047 0.0	
3417302	2	1539548	1540819	putative nitrate/nitrite transporter; NarD [Pseudomonas fluorescens]	gb|AAG34371.1|	1E-107	Nitrogen Metabolism
3174302	2	1540847	1541641	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE C [Brucella	ref|NP_541932.1|	127 2e-28	

				melitensis]			
3174502	2	1542515	1543201	nitrate reductase 1 delta subunit [Shigella flexneri 2a str. 301]	ref|NP_707135.1|	132 5e-30	Operon (nitrogen)
2847802	2	1543201	1544820	respiratory nitrate reductase, beta subunit [Brucella suis 1330]	ref|NP_699493.1|	786 0.0	Operon (nitrogen)
2847902	2	1544727	1548527	probable nitrate reductase alpha subunit [Halomonas halodenitrificans]	dbj|BAB84312.1|	1806 0.0	Operon (nitrogen)
2848002	2	1562125	1564590	NosB [Nostoc sp. GSV224] Polyketide synthase Type I	gb|AAF15892.2|	400 e-110	
2880502	2	1609219	1610358	SyrP-like protein [Streptomyces verticillus]	gb|AAG02366.1|	263 3e-69	
3225002	2	1878953	1879795	PROBABLE GLUTATHIONE S-TRANSFERASE PROTEIN [Ralstonia solanacearum]	ref|NP_521191.1	276 2e-73	
3426202	2	2466228	2467268	putative 4-hydroxy-2-oxovalerate aldolase (EC 4.1.3.-) [Streptomyces coelicolor A3(2)]	ref|NP_639628.1|	288 9e-77	
3199602	2	2469956	2470636	dlpA protein - Legionella pneumophila	sp|Q48806|	89 5e-17	Operon (Barbamide)
3149002	2	2681882	2682931	deoxyribonuclease II [Rattus norvegicus]	ref|NP_612548.1|	147 2e-34	
3414002	2	2701467	2702885	beta-lactamase [Burkholderia pseudomallei]	emb|CAD32564.1|	548 e-155	Drug Resistance
3414102	2	2708927	2709646	Glutamate/aspartate Periplasmic binding protein precursor [Escherichia coli CFT073]	ref|NP_752669.1|	180 2e-44	
3444802	2	2731213	2732691	CG8946-PA [Drosophila melanogaster]	ref|NP_652032.1|	323 5e-87	
3248702	2	2754806	2755738	PUTATIVE CYCLOPROPANE-MYCOLIC ACID SYNTHASE (METHYLTRANSFERASE) PROTEIN [Ralstonia solanacearum]	ref|NP_523216.1|	137 3e-31	
2854602	2	2816331	2818637	putative alpha-D-galactosidase [Yersinia pestis KIM]	ref|NP_669886.1|	526 e-148	
3428402	2	2819359	2820309	putative ABC transporter, permease [Sinorhizobium meliloti]	ref|NP_436488.1|	291 9e-78	Deleted in *B. thailandensis*
3009602	2	2941731	2943116	putative epoxide hydrolase protein [Sinorhizobium meliloti]	ref|NP_436749.1|	442 e-123	
2808902	2	2967211	2968782	tyrosine aminotransferase, tyrosine repressible [Escherichia coli K12]	ref|NP_418478.1|	459 e-128	
				Deleted in *B. thailandensis*			
3218202	1	369368	370294	PROBABLE DIHYDRODIPICOLINATE SYNTHASE-RELATED TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_518786.1	200 3e-50	Deleted in *B. mallei*
2808202	1	516797	518146	RhbA diaminobutyrate-pyruvate aminotransferase [Sinorhizobium meliloti]	ref|NP_436504.1|	453 e-126	
3415502	1	1229074	1229997	probable short-chain dehydrogenase [Pseudomonas aeruginosa PA01]	ref|NP_252014.1|	180 3e-44	
3298702	1	1958241	1959839	Urocanase [Bacillus anthracis A2012]	ref|NP_657554.1|	511 e-143	
3295502	1	1994026	1996704	penicillin acylase II [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_636766.1|	466 e-130	
3000402	1	2030614	2031765	NikT protein [Streptomyces tendae]	emb|CAC11142.1|	102 1e-20	Nikkomycin Biosynthesis
3132802	1	2413643	2415478	EMRY_ECOLI MULTIDRUG RESISTANCE PROTEIN Y	sp|P52600	320 5e-86	Drug
2874802	1	3198568	3199623	putative undecaprenyl phosphate N-acetylglucosaminyltransferase WbiH [Burkholderia pseudomallei]	gb|AAD05467.1|	540 e-152	

2874502	1	3201463	3203331	putative glycosyl transferase [Burkholderia mallei]	gb|AAK27401.1|	0	
3284702	1	3229270	3229932	cinnamoyl ester hydrolase [Bacillus halodurans]	ref|NP_241603.1|	138 7e-32	
2965502	1	4012435	4012965	putative oxidoreductase protein [Sinorhizobium meliloti]	ref|NP_437248.1|	85 1e-26	
3413302	2	266543	268738	penicillin acylase (penicillin amidase) [Oceanobacillus iheyensis]	ref|NP_694226.1|	485 e-135	
3361402	2	831177	833339	putative phosphoenolpyruvate decarboxylase [Bacteroides fragilis]	gb|AAG26466.1|	329 1e-88	Deleted in *B. mallei*
2937602	2	1042018	1043697	PROBABLE VANILLIN DEHYDROGENASE OXIDOREDUCTASE PROTEIN [Ralstonia solanacearum]	ref|NP_521787.1|	460 e-128	
2793202	2	1043679	1045532	PROBABLE FERULOYL-COA SYNTHETASE PROTEIN [Ralstonia solanacearum]	ref|NP_521788.1|	712 0.0	
3333302	2	1111237	1112454	PUTATIVE ISOPENICILLIN N EPIMERASE PROTEIN [Ralstonia solanacearum]	ref|NP_522257.1	191 1e-47	Drug
2829202	2	1322015	1323271	TaC protein [Myxococcus xanthus]	emb|CAB46502.1|	567 e-160	Polyketide biosynthesis
2878102	2	1355303	1357003	polyketide synthase [Bacillus subtilis]	ref|NP_389604.1|	276 9e-73	Polyketide biosynthesis
3014302	2	1487422	1488204	putative S-adenosyl-methionine-sterol-C-methyltransferase, 5' partial; 1-1344 [Arabidopsis thaliana]	gb|AAG52075.1|	150 1e-35	
3014102	2	1488211	1489524	epsilon-trimethyllysine hydroxylase [Rattus norvegicus]	ref|NP_596878.1	89 1e-16	Deleted in *B. mallei*
2901002	2	1708188	1710374	ToxR-regulated lipoprotein [Escherichia coli]	ref|NP_052607.1	84 1e-14	
2896002	2	1718979	1720223	non-ribosomal peptide synthetase [Streptomyces avermitilis]	dbj|BAB69380.1|	169 1e-40	Polyketide biosynthesis
3314802	2	1721311	1726371	epoD [Polyangium cellulosum]	gb|AAF62883.1	296 3e-78	Polyketide biosynthesis
2923702	2	2184502	2185422	AGR_L_4p [Agrobacterium tumefaciens] ref|NP_535374.1| acetyltransferase [Agrobacterium tumefaciens str. C58 (U. Washington)]	ref|NP_355763.1|	0.0000002	
3431502	2	2206195	2207157	SyrP-like protein [Streptomyces avermitilis]	dbj|BAB69339.1	216 4e-55	
3214402	2	2253464	2257933	putative 4-hydroxybutyrate coenzyme A transferase [Yersinia pestis]	ref|NP_405485.1	514 e-144	
3428402	2	2819359	2820309	putative ABC transporter, permease [Sinorhizobium meliloti]	ref|NP_436488.1|	291 9e-78	Deleted in *B. mallei*
3376302	2	2859369	2860652	probable FAD-dependent monooxygenase [Pseudomonas aeruginosa PA01]	ref|NP_252907.1|	237 3e-61	
3365602	2	3023644	3024897	nitric oxide reductase regulator NorR2	emb|CAC00712.1	208 1e-52	
3365802	2	3028187	3029938	putative transferase [Streptomyces coelicolor A3(2)]	emb|CAD55374.1|	456 e-127	
2889502	2	3104374	3105864	disulphide-isomerase [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_644184.1|	244 3e-63	
				TRANSCRIPTION Deleted in *B. mallei*			

3432902	1	362187	362642	PROBABLE TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_522936.1|	251 2e-66	
3301102	1	1915384	1916205	AGR_L_1259p [Agrobacterium tumefaciens] pir||E98210 transcription regulator, gntR family [imported]	ref|NP_356422.1|	8E-45	
3300702	1	1921268	1921720	PUTATIVE HNS-LIKE TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_521590.1|	66 1e-10	Deleted in *B. thailandensis*
2790202	2	794723	795775	transcriptional regulator PchR [Pseudomonas aeruginosa PA01]	ref|NP_252917.1|	404 e-111	
3003802	2	2976263	2977174	PROBABLE TRANSCRIPTIONAL REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_523130.1|	211 1e-53	
3417602	2	1552886	1553662	PROBABLE TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_518311.1	246 2e-64	
3003402	2	2966632	2967249	probable transcriptional regulator [Pseudomonas aeruginosa PA01]	ref|NP_250772.1|	188 5e-47	
				Deleted in *B. thailandensis*			
2957202	1	536736	538097	probable transcriptional regulator [Pseudomonas aeruginosa PA01]	ref|NP_253905.1|	243 5e-63	
3313902	1	1528813	1530351	probable transcriptional regulator [Mesorhizobium loti]	ref|NP_103317.1	254 2e-66	
3300702	1	1921268	1921720	PUTATIVE HNS-LIKE TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_521590.1|	66 1e-10	Deleted in *B. mallei*
2978802	1	2475152	2476216	transcriptional regulator, LysR family [Pseudomonas putida KT2440]	ref|NP_745768.1	77 4e-13	
3170102	1	3305700	3306125	PROBABLE FLAGELLAR TRANSCRIPTIONAL ACTIVATOR TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_522972.1|	0.000001	
3234402	1	3654384	3655325	transcriptional regulator, AraC family [Caulobacter crescentus CB15]	ref|NP_422300.1|	67 3e-10	
3282302	2	56075	56704	RNA polymerase ECF-type sigma factor [Mesorhizobium loti]	ref|NP_108303.1|	87 2e-16	
3413202	2	264490	265506	probable transcriptional regulator [Pseudomonas aeruginosa PA01]	ref|NP_249182.1|	379 e-104	
				MEMBRANE			
				Deleted in *B. mallei*			
3054002	1	1447046	1448860	PROBABLE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_521693.1|	243 9e-63	
3180502	1	3714079	3714699	PUTATIVE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_522302.1|	129 3e-29	
3126302	2	107264	107752	PUTATIVE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_521722.1|	169 1e-41	
3126002	2	110182	110724	putative fimbrial protein [Yersinia pestis KIM]	ref|NP_668859.1|	103 2e-21	Operon
3125902	2	114334	114948	F17A fimbrial subunit precursor [Escherichia coli]	gb|AAD17509.1	91 8e-18	Operon
2840602	2	111567	114383	outer membrane usher protein precursor [Xylella fastidiosa 9a5c]	ref|NP_297374.1|	0	
3406102	2	122487	123803	putative membrane protein [Salmonella enterica subsp. enterica serovar Typhi]	ref|NP_454885.1|	361 1e-98	
3421302	2	153170	155734	putative outer membrane fimbrial usher protein [Yersinia pestis]	ref|NP_403954.1|	645 0.0	

3188202	2	175700	176314	PROBABLE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_521264.1|	223 1e-57	
3018702	2	794178	794639	fimbriae-associated protein Fap1 - Streptococcus parasanguinis	pir||T17451	67 1e-10	Deleted in *B. thailandensis*
3343002	2	849813	851192	outer membrane efflux family protein [Shewanella oneidensis MR-1]	ref|NP_716453.1|	5E-91	
3015202	2	1475889	1476479	PROBABLE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_518319.1|	161 5e-39	
3263502	2	2376040	2377191	PROBABLE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_522173.1|	136 5e-31	
3374902	2	2914043	2915524	similar to transporter homolog ywoD [Bacillus subtilis]	ref|NP_391529.1|	200 4e-50	
3353402	2	2974830	2976134	PUTATIVE TRANSPORT TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_523131.1|	417 e-115	
				Deleted in *B. thailandensis*			
2798902	1	2141246	2143711	MrfC [Photorhabdus luminescens]	gb|AAK82425.1|	535 e-150	Fimbrae component
3470002	1	1885085	1887634	putative outer membrane protein, export function [Escherichia coli O157:H7 EDL933]	ref|NP_286281.1|	513 e-144	
3126102	2	110164	110874	frimbrial protein [Yersinia pestis]	ref|NP_406441.1|	102 4e-21	Operon
3186802	2	155915	156622	putative fimbrial chaperone [Yersinia pestis]	ref|NP_403955.1|	197 8e-50	Operon
3018702	2	794178	794638	fimbriae-associated protein Fap1 - Streptococcus parasanguinis	pir||T17451	67 1e-10	Deleted in B. mallei
3192302	2	965130	966395	PROBABLE TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_519103.1|	275 8e-73	
3435002	2	1883056	1884663	PUTATIVE TRANSMEMBRANE CYTOCHROME BD-II OXIDASE (SUBUNIT I) OXIDOREDUCTASE [Ralstonia solanacearum]	ref|NP_522768.1|	534 e-150	
2954302	2	1609219	1610358	putative transport protein [Salmonella typhimurium LT2]	ref|NP_460503.1|	412 e-114	
3020602	2	2033784	2035331	PUTATIVE HEMAGGLUTININ-RELATED PROTEIN [Ralstonia solanacearum]	ref|NP_523179.1|	184 2e-45	
3450802	2	2370880	2372571	FhaC [Bordetella bronchiseptica]	gb|AAF21946.1|AF111795_1	282 1e-74	Outer membrane protein
				SIGNALING			
				Deleted in *B. mallei*			
2858102	1	359490	361643	non-hemolytic phospholipase C [Burkholderia thailandensis]	gb|AAM55475.1|	1347 0.0	
3302202	1	1899078	1900508	GGDEF family protein [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_642264.1|	135 1e-30	Deleted in *B. thailandensis*
3176002	2	1584149	1584826	PROBABLE TRANSCRIPTIONAL ACTIVATOR OF QUORUM SENSING [Ralstonia solanacearum]	ref|NP_522339.1|	181 6e-45	Quorum sensing
3418002	2	1588194	1588811	acyl-homoserine lactone synthase BviI [Burkholderia cepacia]	gb|AAK35155.1|	187 1e-46	Quorum sensing
				Deleted in *B. thailandensis*			

2912602	1	1512096	1512503	HYPOTHETICAL SIGNAL PEPTIDE PROTEIN [Ralstonia solanacearum]	ref|NP_521676.1|	132 1e-30	
3302202	1	1899078	1900508	GGDEF family protein [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_642264.1|	135 1e-30	Often found in associated with sensor components
3299702	1	1948084	1948794	PROBABLE TWO-COMPONENT SYSTEM RESPONSE REGULATOR TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_519194.1|	146 2e-34	
2872602	2	69372	70787	PROBABLE RESPONSE REGULATOR OF THE C4-DICARBOXYLATE TRANSPORT TWO-COMPONENT REGULATORY SYSTEM TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_521570.1|	569 e-161	
3421202	2	151959	152618	probable response regulator [Pseudomonas aeruginosa PA01]	ref|NP_252769.1|	199 2e-50	
3161502	2	898946	899653	two-component response regulatory protein [Salmonella enterica subsp. enterica serovar Typhi]	ref|NP_457992.1|	80 3e-14	
2897002	2	1047507	1048643	PROBABLE PORIN SIGNAL PEPTIDE PROTEIN [Ralstonia solanacearum]	ref|NP_523153.1|	347 2e-94	
				SECRETED			
				Deleted in *B. mallei*			
3445702	2	2770769	2780077	hemagglutinin-like secreted protein XF2196 [imported] - Xylella fastidiosa (strain 9a5c)	pir||E82589	392 e-107	
				Deleted in B. thailandensis			
3300202	1	1936238	1938637	toxin secretion ATP-binding protein VCA1084 [imported] - Vibrio cholerae (strain N16961 serogroup O1)	pir||H82381	334 3e-90	
3300102	1	1938541	1940043	secretion protein, HlyD family VCA1080 [imported] - Vibrio cholerae (strain N16961 serogroup O1)	pir||D82381	225 5e-58	
3187002	2	157971	158666	putative exported protein [Yersinia pestis]	ref|NP_403953.1|	183 2e-45	Deleted in *B. mallei*

Table S12: ORFs deleted in *B. mallei* and *B. thailandensis* compared to *B. pseudomallei*. Not shown are I) ORFs with no homology or with homology to hypothetical proteins, II) ORFs which have been previously identified as being differentially present among the three species, III) ORFs with homology to insertion sequences, transposases, and other mobilizable elements.

Supplementary Data S13: Deleted ORFs Belonging to RDs ("Regions of Difference") 1-12

NCCID	Chr	Start	Stop	Homology	Accession	RD Number
3418702	1	92521	93762	hypothetical protein [Burkholderia fungorum]	gb|ZP_00028268.1|	RD1
3179402	1	94416	97280	putative inner membrane protein [Salmonella typhimurium LT2]	ref|NP_461672.1|	
3179602	1	97746	98540	NA	NA	
3179802	1	99412	100212	NA	NA	
3180002	1	101251	102399	HYPOTHETICAL PROTEIN [Ralstonia solanacearum]	ref|NP_521827.1|	
3134902	1	147391	150180	HYPOTHETICAL PROTEIN [Ralstonia solanacearum]	ref|NP_519088.1|	RD2
3134802	1	150395	150892	NA	NA	
3134702	1	151566	152042	NA	NA	
3134202	1	155399	156493	phage-related tail protein [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_638342.1|	
3134102	1	156496	156918	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00089325.1|	
3133902	1	159844	160938	phage-related tail protein [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_638337.1|	
3133802	1	161970	162707	PROBABLE PHAGE-RELATED PROTEIN (TAIL SHEATH-LIKE PROTEIN) [Ralstonia solanacearum]	ref|NP_520039.1|	
3133702	1	162846	164218	hypothetical protein [Pseudomonas fluorescens]	gb|ZP_00084026.1|	
3133602	1	162867	164135	hypothetical protein [Pseudomonas fluorescens]	gb|ZP_00084026.1|	
3050602	1	165241	165792	phage-related tail protein [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_642967.1|	
3050502	1	165788	166954	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088456.1|	
3050402	1	166250	166810	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088456.1|	
3050302	1	167049	167798	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088458.1|	
3050202	1	167734	168744	site-specific DNA-methyltransferase [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_638315.1|	
3050102	1	169190	169711	orf14; similar to S gene of P2:tail completion [Bacteriophage phi CTX]	ref|NP_490613.1|	
3050002	1	169525	169938	orf13; similar to R gene of P2:tail completion [Bacteriophage phi CTX]	ref|NP_490612.1|	
3049902	1	170046	170483	NA	NA	
3372302	1	170483	171292	PUTATIVE PHAGE-RELATED PROTEIN (HYDROLASE) [Ralstonia solanacearum]	ref|NP_520052.1|	
3049802	1	172379	172894	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088466.1|	
3049702	1	172958	173683	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088467.1|	
3049502	1	173644	174651	orf4; similar to N of P2:major capsid protein [Bacteriophage phi CTX]	ref|NP_490602.1|	
3049602	1	174607	175839	PROBABLE BACTERIOPHAGE PROTEIN [Ralstonia solanacearum]	ref|NP_520059.1|	
3049402	1	174688	175533	PROBABLE BACTERIOPHAGE PROTEIN [Ralstonia solanacearum]	ref|NP_520059.1|	
3049302	1	175842	177404	orf2; similar to P gene of P2 :DNA-dependent ATPase terminase subunit [Bacteriophage phi CTX]	ref|NP_490600.1|	
3269502	1	847846	849039	hypothetical protein [Burkholderia fungorum]	gb|ZP_00034165.1|	RD3

3269302	1	850345	850878	NA	NA	
3269202	1	851788	852963	NA	NA	
3269102	1	853004	854032	AGR_pAT_24p [Agrobacterium tumefaciens]	ref|NP_395955.1|	
3268902	1	854496	855059	NA	NA	
3268802	1	855097	855651	HYPOTHETICAL/UNKNOWN PROTEIN [Ralstonia solanacearum]	ref|NP_520732.1|	
3268702	1	855650	856558	HYPOTHETICAL TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_520733.1|	
3268602	1	855656	856495	HYPOTHETICAL TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_520733.1|	
3268502	1	856589	857374	hypothetical protein [Vibrio cholerae]	ref|NP_232790.1|	
3452602	1	856595	857440	hypothetical protein [Vibrio cholerae]	ref|NP_232790.1|	
3268402	1	857477	858832	integrase [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_643606.1|	
3302902	1	1893743	1894492	NA	NA	RD4
3302702	1	1894351	1895034	NA	NA	
3302602	1	1894363	1894974	NA	NA	
3302402	1	1895264	1895926	hypothetical protein [Ralstonia metallidurans]	gb|ZP_00024986.1|	
3302302	1	1895929	1899174	hypothetical protein [Ralstonia metallidurans]	gb|ZP_00024985.1|	
3302202	1	1899078	1900508	GGDEF family protein [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_642264.1|	
3302002	1	1900778	1901584	NA	NA	
3301802	1	1904329	1904778	ISRSO7-TRANSPOSASE PROTEIN [Ralstonia solanacearum]	ref|NP_518373.1|	
3301702	1	1904893	1905732	transcriptional regulator, GntR family [Agrobacterium tumefaciens str. C58 (U. Washington)]	ref|NP_534710.1|	
3301602	1	1905946	1906566	hypothetical protein with local similarity [Sinorhizobium meliloti]	ref|NP_435673.1|	
3301502	1	1905994	1906617	hypothetical protein with local similarity [Sinorhizobium meliloti]	ref|NP_435673.1|	
3301402	1	1907457	1908440	AGR_pAT_322p [Agrobacterium tumefaciens]	ref|NP_396159.1|	
3469902	1	1908552	1909592	conserved hypothetical protein [Sinorhizobium meliloti]	ref|NP_435676.1|	
3301302	1	1912663	1913556	putative ABC transporter, permease protein [Sinorhizobium meliloti]	ref|NP_435680.1|	
3301202	1	1913440	1914249	putative ABC transporter, permease protein [Sinorhizobium meliloti]	ref|NP_435681.1|	
3469602	1	1914372	1915460	putative ABC-transporter ATP-binding protein [Sinorhizobium meliloti]	ref|NP_435682.1|	
3469502	1	1916230	1917687	hypothetical protein [Burkholderia fungorum]	gb|ZP_00032798.1|	
3300902	1	1917777	1918949	hypothetical protein [Burkholderia fungorum]	gb|ZP_00031010.1|	
3300802	1	1918957	1919364	NA	NA	
3300702	1	1921268	1921720	PUTATIVE HNS-LIKE TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_521590.1|	
3300602	1	1922210	1923106	hypothetical protein [Burkholderia fungorum]	gb|ZP_00028591.1|	
3300502	1	1923106	1925034	hypothetical protein [Magnetococcus sp. MC-1]	gb|ZP_00043031.1|	
3300402	1	1934933	1935439	NA	NA	
3300302	1	1935424	1936311	hypothetical protein [Pseudomonas fluorescens]	gb|ZP_00086357.1|	
3300202	1	1936238	1938637	hypothetical protein [Desulfovibrio desulfuricans G20]	gb|ZP_00130819.1|	
3300102	1	1938541	1940043	hypothetical protein [Pseudomonas fluorescens]	gb|ZP_00086337.1|	

3300002	1	1940157	1944518	hypothetical protein [Trichodesmium erythraeum IMS101]	gb|ZP_00072341.1|	
3299902	1	1944491	1946185	HYPOTHETICAL PROTEIN [Ralstonia solanacearum]	ref|NP_521285.1|	
3299702	1	1948084	1948794	PROBABLE TWO-COMPONENT SYSTEM RESPONSE REGULATOR TRANSCRIPTION REGULATOR PROTEIN [Ralstonia solanacearum]	ref|NP_519194.1|	
3294402	1	3097526	3098038	NA	NA	RD5
3294202	1	3098307	3098876	hypothetical protein [Burkholderia fungorum]	gb|ZP_00030699.1|	
3294102	1	3098814	3099368	NA	NA	
3294002	1	3099384	3100373	NA	NA	
3293902	1	3101272	3101904	NA	NA	
3180102	1	3718434	3719231	NA	NA	RD6
3418802	1	3719241	3720308	transcriptional regulator, HTH_3 family [Neisseria meningitidis MC58]	ref|NP_275004.1|	
3026702	1	3721641	3722048	hypothetical protein [Pseudomonas syringae pv. syringae B728a]	gb|ZP_00127047.1|	
3026802	1	3722120	3724717	type I restriction-modification system, M subunit, putative [Shewanella oneidensis MR-1]	ref|NP_717074.1|	
3283102	2	78855	79598	hypothetical protein [Pseudomonas aeruginosa PA01]	ref|NP_250201.1|	RD7
2872802	2	79601	80011	ISBt3 transposase subunit protein [Bacteriophage phiE125]	ref|NP_536395.1|	
2872902	2	81347	82069	hypothetical protein [Ralstonia metallidurans]	gb|ZP_00025020.1|	
3457402	2	82069	82485	Insertion element IS401 hypothetical 12.4 kDa protein gb|AAA25861.1| IS401 transposase subunit; putative [Burkholderia cepacia]	sp|Q51647|YISX_BURCE	
3283202	2	82941	84035	probable transcriptional regulator [Pseudomonas aeruginosa PA01]	ref|NP_251394.1|	
3283302	2	83980	84564	hypothetical protein [Rhodospirillum rubrum]	gb|ZP_00015144.1|	
3457502	2	84913	85830	probable permease of ABC transporter [Pseudomonas aeruginosa PA01]	ref|NP_249016.1|	
3457602	2	85817	86668	probable permease of ABC transporter [Pseudomonas aeruginosa PA01]	ref|NP_249015.1|	
3457702	2	86668	87696	probable binding protein component of ABC transporter [Pseudomonas aeruginosa PA01]	ref|NP_249014.1|	
3127302	2	93965	95611	hypothetical protein [Pseudomonas syringae pv. syringae B728a]	gb|ZP_00127913.1|	
3273102	2	539338	539742	conserved hypothetical protein [Xylella fastidiosa 9a5c]	ref|NP_298851.1|	RD8
3273202	2	540575	541036	hypothetical protein [Bacteriophage phiE125]	ref|NP_536405.1|	
3273302	2	541540	542325	NA	NA	
3273402	2	542310	543695	conserved hypothetical protein [Bacteriophage phiE125]	ref|NP_536400.1|	
3273502	2	544098	544748	putative truncated transposase [Xanthomonas campestris pv. vesicatoria]	gb|AAL78293.1|	
3191302	2	978884	979480	hypothetical protein [Burkholderia fungorum]	gb|ZP_00029497.1|	RD9
3191002	2	982464	983273	hypothetical protein [Pseudomonas aeruginosa PA01]	ref|NP_249672.1|	
3190902	2	982533	983228	hypothetical protein [Pseudomonas aeruginosa PA01]	ref|NP_249672.1|	
3157702	2	1437139	1437657	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00092036.1|	RD10
3158202	2	1442842	1443567	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088467.1|	
3158302	2	1443631	1444146	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00088466.1|	
3158402	2	1445233	1446042	PUTATIVE PHAGE-RELATED PROTEIN (HYDROLASE) [Ralstonia solanacearum]	ref|NP_520052.1|	

3158602	2	1447594	1448121	NA	NA	
3398202	2	1448293	1449165	site-specific DNA-methyltransferase [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_638315.1|	
3158802	2	1451174	1451725	phage-related tail protein [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_642967.1|	
3158902	2	1454119	1454787	PUTATIVE TAIL FIBER ASSEMBLY-LIKE PROTEIN [Ralstonia solanacearum]	ref|NP_520041.1|	
2841502	2	1454927	1456015	PROBABLE PHAGE-RELATED PROTEIN (TAIL SHEATH-LIKE PROTEIN) [Ralstonia solanacearum]	ref|NP_520039.1|	
3628902	2	1457064	1460027	PROBABLE PHAGE-RELATED TAIL TRANSMEMBRANE PROTEIN [Ralstonia solanacearum]	ref|NP_520035.1|	
3629002	2	1460048	1460470	hypothetical protein [Azotobacter vinelandii]	gb|ZP_00089325.1|	
4028002	2	1460470	1461570	phage-related tail protein [Xanthomonas campestris pv. campestris str. ATCC 33913]	ref|NP_638342.1|	
4146202	2	2761507	2762169	conserved hypothetical protein [Xanthomonas axonopodis pv. citri str. 306]	ref|NP_643040.1|	RD11
3950202	2	2763240	2764787	Cytosine deaminase and related metal-dependent hydrolases [Thermoanaerobacter tengcongensis]	ref|NP_623193.1|	
3950102	2	2763375	2763908	PUTATIVE HYDROLASE PROTEIN [Sinorhizobium meliloti]	ref|NP_386696.1|	
3950302	2	2764618	2765538	glutathione S-transferase family protein [Caulobacter crescentus CB15]	ref|NP_421446.1|	
4146302	2	2764813	2765466	glutathione S-transferase family protein [Caulobacter crescentus CB15]	ref|NP_421446.1|	
3950602	2	2767013	2768149	PUTATIVE TRANSPOSASE PROTEIN [Ralstonia solanacearum]	ref|NP_521300.1|	
4146402	2	2768152	2769927	hypothetical protein [Burkholderia fungorum]	gb|ZP_00033628.1|	
4146502	2	2770769	2780077	hypothetical protein [Burkholderia fungorum]	gb|ZP_00031982.1|	
3951002	2	2785931	2787196	hypothetical [Yersinia pestis KIM]	ref|NP_670729.1|	
3758902	2	2909989	2910858	putative citrate lyase beta chain [Streptomyces coelicolor A3(2)]	ref|NP_626293.1|	RD12
3519502	2	2910858	2912156	hypothetical protein [Streptomyces coelicolor A3(2)]	ref|NP_626292.1|	
3374802	2	2913402	2914043	putative hydrolase [Streptomyces coelicolor A3(2)]	ref|NP_625051.1|	
3058202	2	2913438	2914037	putative hydrolase [Streptomyces coelicolor A3(2)]	ref|NP_625051.1|	
3374902	2	2914043	2915524	similar to transporter [Bacillus subtilis]	ref|NP_391529.1|	
3058302	2	2914058	2915485	similar to transporter [Bacillus subtilis]	ref|NP_391529.1|	
3058402	2	2917576	2918610	hypothetical protein [Ralstonia metallidurans]	gb|ZP_00021404.1|	
3375102	2	2918648	2919922	hypothetical protein [Streptomyces coelicolor A3(2)]	ref|NP_626292.1|	

Table S13: ORFs belonging to RD1-RD12, defined by molecular subtyping of *B. pseudomallei* strains.

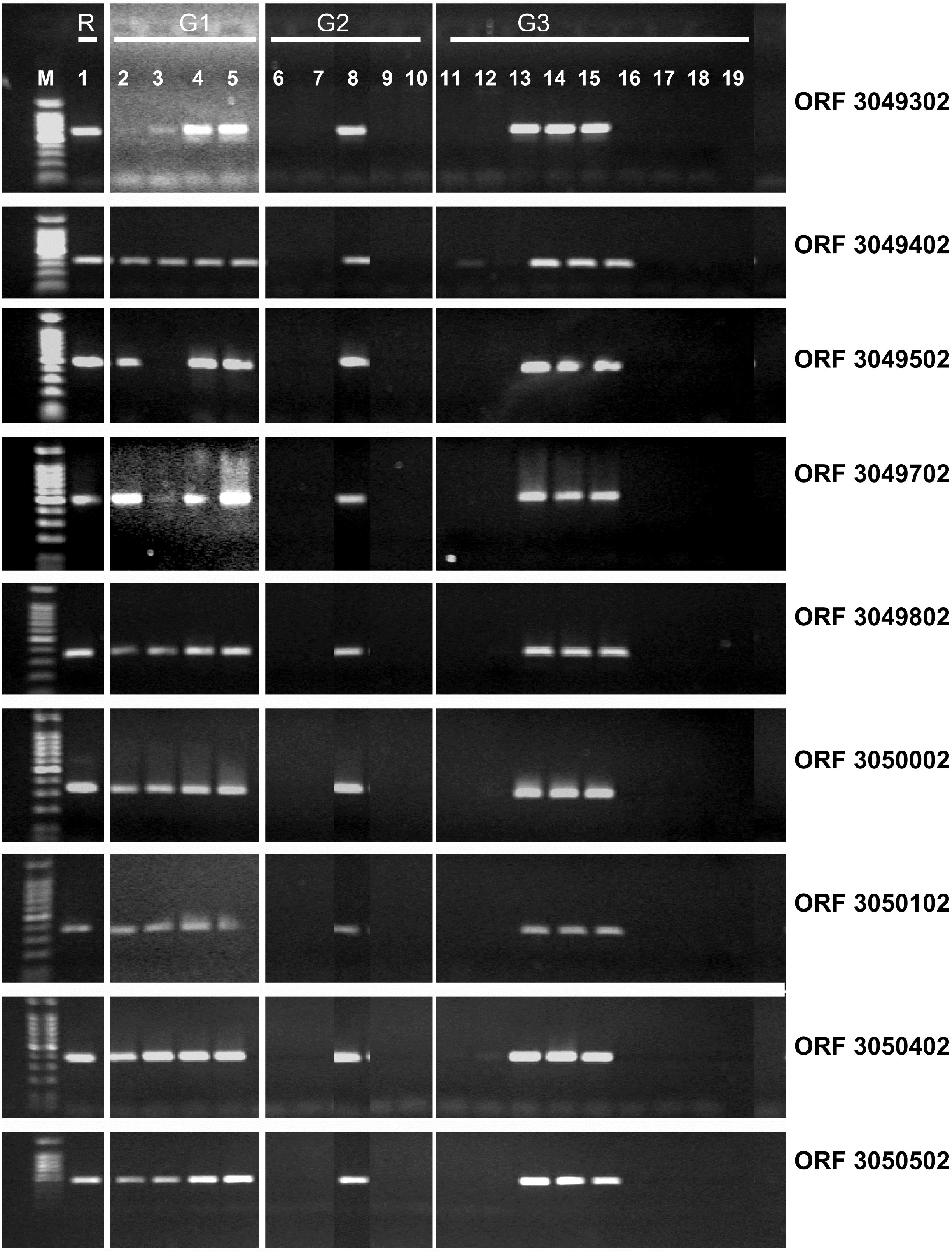
Supplementary Data S14: Relationships of RD1-RD12 Among *B. pseudomallei* molecular subtypes

RD	Chr	Pattern	Description
RD01	1	P1	Deleted in G3
RD02	1	P3	Deleted in G2 and G3
RD03	1	P3	Deleted in G2 and G3
RD04	1	P5	Amplified in G1
RD05	1	P2	Amplified in G1, Deleted in G3
RD06	1	P4	Deleted in G1 and G2
RD07	2	P4	Deleted in G1 and G2
RD08	2	P1	Deleted in G3
RD09	2	P3	Deleted in G2 and G3
RD10	2	P3	Deleted in G2 and G3
RD11	2	Mixture	Mixture of G1 deletion and G2/G3 deletions
RD12	2	P4	Deleted in G1 and G2

Supplementary Data S15: Deletion of the RD2 CTX-like Cluster in *B. pseudomallei* Isolates.

Lanes 1 – 19: *B. pseudomallei* strains - K96243 (reference R), G1 subtype - 5/96, 567/96, 21/96, 497/96, G2 subtype - #33, #22, #11, #23, #14, G3 subtype – 20/96, 15/96, ATCC23343, #59, #9, 78/96, 15-10, ATCC15682, #3. Lanes have been arranged to be similar to Figure 5 in the Main Text. M – 100bp molecular marker. ORF primer pairs are homologous to the following CTX proteins: 3049302 (orf 2, DNA-dependent ATPase terminase subunit), 3049402 (orf 3, presumed capsid scaffold), 3049502 (orf 4, major capsid protein), 3049702 (orf 5, terminase subunit), 3049802 (head completion/stabilization protein (gene L) of phage P2), 3050002 (orf 13, tail completion), 3050102 (orf 14, tail completion), 3050402 (orf 18, baseplate or base of tail fiber), 3050502 (orf 18, baseplate or base of tail fiber).

Figure S15: PCR Validation of RD2 In *B. pseudomallei* Isolates



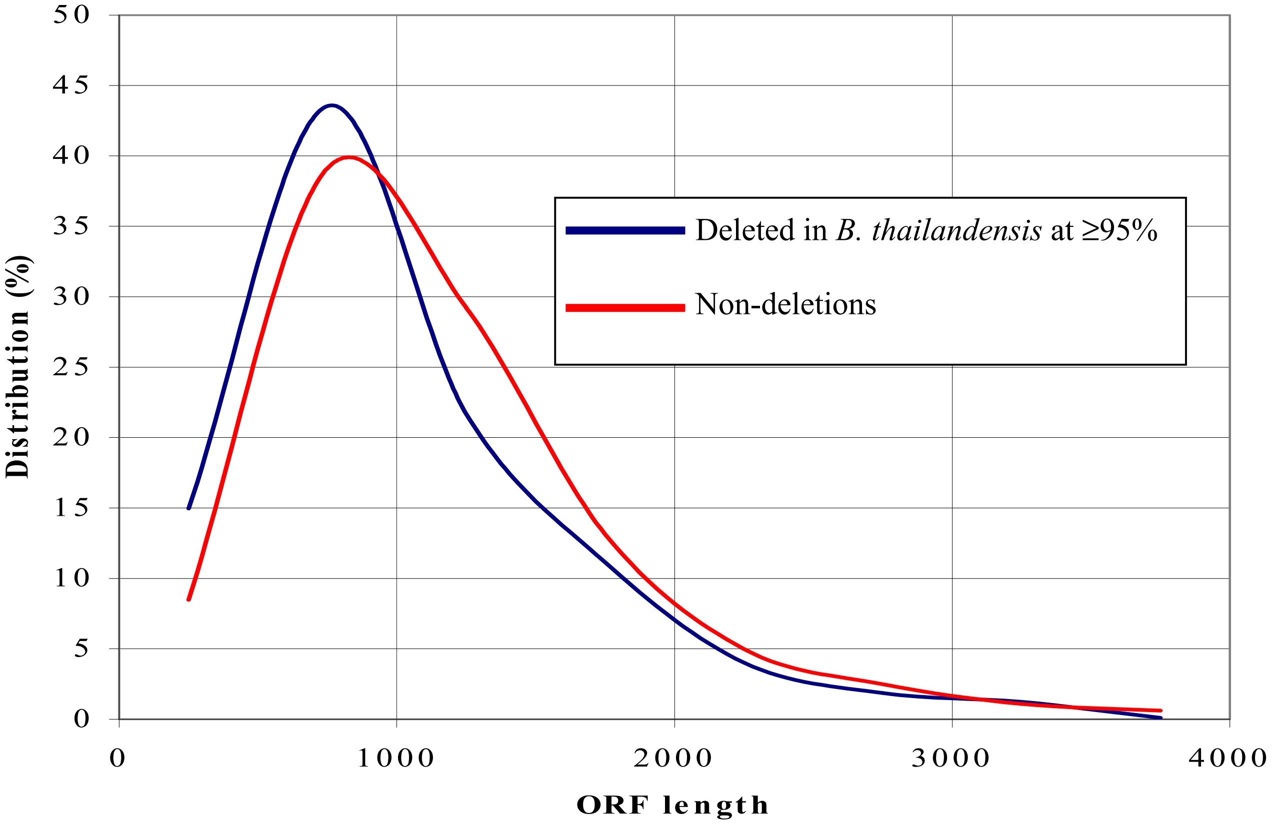
Supplementary Data S16: Comparison of ORF Length and Codon Usages Between "Deleted" and Non-deleted ORFs in *B. thailandensis*

The 6895 ORFs were first divided into 2 groups:

(a) Group 1: ORFs analysed as deleted in *B. thailandensis* at ≥95% level of confidence (969 ORFs)
(b) Group 2: The rest of the ORFs (5926 ORFs) (Non-deletions)

(A) ORF length

Figure S16a: ORF-length-based Distribution for Deleted ORFs and Non-Deletions



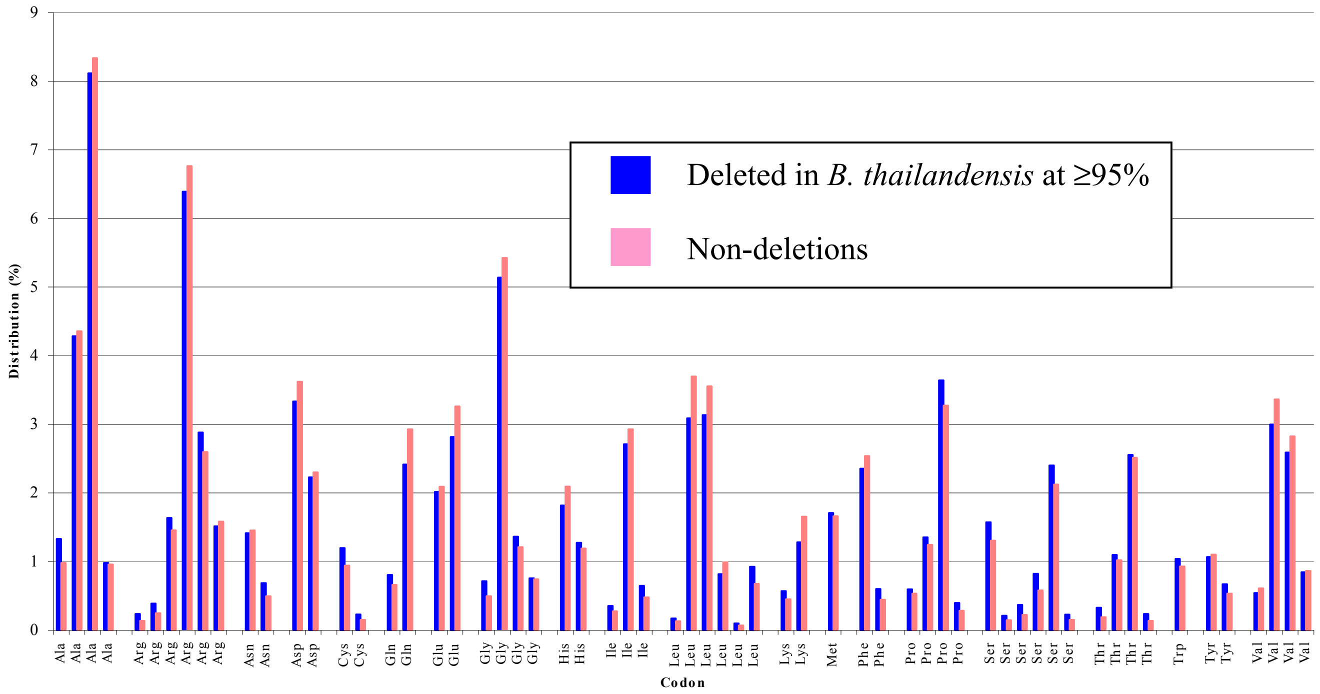
Summary of ORF-length comparison between the 2 groups of ORFs

	Median	Mean	Mod range
Group 1	888	1125.75	400~600
Group 2	1029	1196.32	400~600

(B) Codon usage frequencies

For codon usage analysis, we collected the frequency of each of the 64 genetic codes for the ORFs in both groups. In order to compare the codon usage statistics between the 2 groups, the raw frequencies obtained for the ORFs of each group were divided by the total number of codons in their respective groups (i.e. 363940 and 2362808 total codons for Group 1 and 2 respectively). These normalized frequencies (in %) were then used to make feasible comparisons between Groups 1 and 2, and are shown in Figure S16b.

Figure S16b: Codon-usage-based Distribution for Deleted ORFs and Non-Deletions



No significance difference between the 2 groups in terms of codon usage frequencies. In fact, the average of differences between Groups 1 and 2 in all codon is negligible (4×10^{-17}) while the standard deviation stands at 0.22%.