Genomes link the global spread of *Burkholderia pseudomallei* to human movement and trade


* Correspondence should be addressed to Claire Chewapreecha (cchewapreecha@gmail.com) and Sharon Peacock (sharon.peacock@lshtm.ac.uk)

1 Department of Medicine, University of Cambridge, UK
2 Wellcome Trust Sanger Institute, Cambridge, UK
3 Systems Biology and Bioinformatics Research Group, King Mongkut’s University of Technology Thonburi, Thailand
4 School of Medicine, University of St Andrew, UK
5 Department of Mathematics and Statistics, University of Helsinki, Finland
6 Department of Medical and Clinical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Finland
7 Department of Veterinary Medicine, University of Cambridge, UK
8 Emerging Pathogens Institute, University of Florida, USA
9 Department of Clinical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
10 Institut Pasteur, Paris, France
Global and Tropical Health Division, Menzies School of Health Research, Charles Darwin University, Darwin, Australia

Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand

Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand

Lao-Oxford-Mahosot Hospital-Wellcome Trust Research Unit, Microbiology Laboratory, Mahosot, Vientiane, Laos PDR

Department of Infectious Disease Epidemiology, Imperial College, UK

Department of Biostatistics, University of Oslo, Oslo, Norway

Center for Microbial Genetics and Genomics, Northern Arizona University, USA

London School of Hygiene and Tropical Medicine, UK
The environmental bacterium *Burkholderia pseudomallei* causes an estimated 165,000 cases of human melioidosis per year worldwide, and is also classified as a biothreat agent. We used whole genome sequences of 469 *B. pseudomallei* isolates from 30 countries collected over 79 years to explore its geographic transmission. Our data point to Australia as an early reservoir, with transmission to Southeast Asia followed by onward transmission to South Asia, and East Asia. Repeated reintroduction was observed within the Malay Peninsula, and between countries bordered by the Mekong river. Our data support an African origin of the Central and South American isolates with introduction of *B. pseudomallei* into the Americas between 1650 and 1850, providing a temporal link with the slave trade. We also identified geographically distinct genes/variants in Australasian or Southeast Asian isolates alone, with virulence-associated genes being among those overrepresented. This provides a potential explanation for clinical manifestations of melioidosis that are geographically restricted.
Burkholderia pseudomallei is an environmental gram-negative bacillus and the cause of melioidosis, a serious disease of humans and animals for which there is no licensed vaccine. Infection results from inoculation, ingestion or inhalation of B. pseudomallei, and is fatal in 10-40% of human cases\(^1\). To further understand the global dissemination of melioidosis, we sequenced 276 B. pseudomallei isolates cultured from humans with melioidosis or from the environment between 1935 and 2013. These originated from 30 countries across Australasia, Asia, Africa and Central and South America. We added to this whole genome data available for a further 193 B. pseudomallei isolates from Southeast Asia\(^2\) and Australia\(^3\), giving a total dataset comprising 469 isolates (See Supplementary table 1 for details of isolates and references). The genetic diversity of these isolates was captured by mapping short-read genome sequences against a core genome created from the two chromosomes of B. pseudomallei K96243\(^4\), and by extracting both core and accessory coding sequences from the assembled genomes (see methods). We employed three different approaches to outline the population structure: phylogenetic reconstructions using single nucleotide polymorphic changes (SNPs) called from core genome mapping (Fig. 1a); SNPs from shared single-copy core genes (Supplementary Fig. 1); and a tree-independent hierarchical Bayesian clustering (Supplementary table 1).

All three approaches demonstrated a clear genetic distinction between isolates from Australasia and Asia (two areas where melioidosis is endemic), supporting previous findings\(^5,6\). Isolates from Australasia had longer phylogenetic branches compared to isolates from other regions, indicative of greater genetic diversity (Figure 1a and Supplementary Fig. 1). This was also observed from the pan-genome analysis\(^7\), which confirmed that the Australasian B. pseudomallei population had the highest rate of new gene discovery and the largest accessory genome (Fig. 1b and 1c).
Examination of data distribution confirmed that this finding was not related to different sampling periods or sequencing platforms used to generate the data (Supplementary Fig. 2). These observations provide evidence for the hypothesis that Australia was an early reservoir for the current global *B. pseudomallei* population\(^5,8\), which is supported by the Australasian isolates being at the base of the tree (Supplementary Fig. 1). An alternative explanation is that there have been repeated population bottlenecks outside Australia, but not within it. Figure 1a and Supplementary Fig. 1 both delineated an apparent single transmission out of Australasia (consistent with previous findings\(^9,10\)), and several independent transmission events from Southeast Asia to South Asia and East Asia. We also noted a monophyly and a single combined Bayesian cluster containing isolates from Africa and Central and South America, suggesting close ancestry (Fig 1a, Supplementary Fig 1 and Supplementary Table 1). The phylogenies also highlighted an African root for this group (100% bootstrap support), implying an African origin of the American isolates based on our sampling density.

We then estimated a timeline for the intercontinental and regional spread of *B. pseudomallei* by identifying and analysing 19 separate Bayesian clusters comprising isolates from Australia and Oceania (group 1), Asia (groups 2 to 18), and Africa and America (group 19). To improve our sensitivity to detect genetic variants, we remapped sequence reads from each cluster against a closely related reference genome (Supplementary Fig 2). After removing sequences that had been horizontally acquired by recombination\(^11\), temporal signals were determined for each cluster with the timeline estimated by BEAST\(^12\) (Supplementary Figs 4, 5 and 6). Clock signals were captured for American isolates within the African-American cluster, and for four Asian clusters. The most recent common ancestor for the American isolates was
estimated to be 1806 and 1759 for chromosome I and II, respectively (combined 95% highest posterior density (HPD) interval of both chromosomes, 1682-1849) (Fig. 2a). The introduction of *B. pseudomallei* into the Americas overlaps with the height of the slave trade between 1650 - 1850, during which an estimated 10-15 million people and related cargoes including environmentally contaminated food and water were transported from Africa to the Americas (Fig. 2b)\(^{13,14}\). Dating of Asian clusters showed that recent common ancestors could be defined for three Malaysian-Singaporean clusters and one Thai – Laos cluster, all of which dated to the 20\(^{th}\) century (Fig. 2a). The most recent common ancestor of other Asian and Australasian clusters is very likely to pre-date these estimates, but dating cannot be reliably assessed using our collection.

Within the Asian isolates, the majority of Southeast Asian clusters either contained isolates from the Malay Peninsula (Malaysia and Singapore – here termed “the Malay sub-region”), or from countries bordered by the Mekong river (Thailand, Laos, Cambodia and Vietnam – here termed “the Mekong sub-region”) (Supplementary Fig. 7a, Supplementary Table 1). To further examine this pattern, we estimated the number of times *B. pseudomallei* transitioned between Southeast Asian countries. This revealed a greater number of transitions within the same sub-regions than between sub-regions (two-tailed Mann-Whitney U test, p-value < 2.2x10\(^{-16}\)) (Supplementary Fig. 7b). The connectivity observed within sub-regions may be explained by geographical proximity, cultural links or trading networks associated with the Mekong river\(^{15,16}\) (Fig. 2c). In addition to an unequal number of transitions, *B. pseudomallei* may have spent different amounts of evolutionary time in these countries (total branch lengths of multiple sub-sampling phylogenetic trees) (Supplementary Fig. 7c). Assuming a homogenous mutation rate, our results are
indicative of a higher proportion of evolutionary time spent in the Mekong versus the Malay sub-region (two-tailed Mann-Whitney U test, p-value $< 2.2 \times 10^{-16}$), and possibly suggests that the Mekong sub-region has been a hotspot for \textit{B. pseudomallei} evolution in the Southeast Asian endemic zone. However, it is possible that the observation may be influenced by evolutionary rate variation on each branch but the local clock cannot be reliably assessed across this dataset (Supplementary Fig. 7).

The most common presentation of human melioidosis in both Asia and Australia is one or more of bacteremia, pneumonia and liver and/or splenic abscesses. By contrast, some of the less common clinical manifestations show geographical segregation, including encephalomyelitis in Australia. Moreover, mortality is lower in Australasia than Southeast Asia (10\% versus 40\%, respectively)\textsuperscript{17}. Differences in access to medical care including intensive care facilities are likely to contribute to different outcomes, but bacterial factors could contribute to disease severity or specific clinical manifestations. To investigate the genetic basis that might explain clinical differences between Australasia and Southeast Asia, we systematically screened for particular kmers (DNA words) that were enriched in Australasian isolates alone, or in Southeast Asian isolates alone using a kmer based GWAS\textsuperscript{18} (see methods). The strong link between the population structure and geographical origin isolates described above led us to omit population stratification in the GWAS analysis (see methods and Supplementary Tables 2, 3 and 4). Kmers were then clustered into loci based on their genetic proximity. This resulted in the identification of 468 and 14 loci that were specific to the Australasian and Southeast Asian population, respectively. Australasia- and Southeast Asia-specific loci were each distributed across multiple phylogenetic branches of their respective population (Supplementary Fig 8), suggesting that these were not solely driven by clonality in the population.
structure but may have been independently acquired and/or lost on multiple occasions. The mechanisms that have driven these patterns are the subject of further investigation.

The region-specific loci detected potentially include variety of loci that may enhance survival and inter-bacterial competition in specific niches, as well as virulence factors that contribute to regionally distinct clinical manifestations. To facilitate the biological interpretation of these data, loci were categorised by the function of genes (COG), gene ontology (GO) and pathway terms. Some genes had no functional match in the curated database, but 64.3% could be assigned and revealed that region-specific genes were widely dispersed across multiple functions (Fig. 3).

Functional enrichment analyses highlighted elevated frequencies of the terms “secondary metabolite biosynthesis”, “translation”, “lipid transport and metabolism” and “defense mechanisms” among region-specific genes compared to random expectation from a reference genome (one-sided Fisher test p-value < 2.2 x 10^{-16}, < 2.2 x 10^{-16}, 1.86 x 10^{-10} and 9.07 x 10^{-10} respectively, Supplementary table 5). The latter contained several virulence genes involved in disease pathogenesis. Our results highlighted several virulence loci with known region specific variations, including *Burkholderia thailandensis*-like flagellum and chemotaxis cluster (BTFC), and *Burkholderia mallei*-like *BimA (BmBimA)*\textsuperscript{19,20}. Both BTFC and *BmBimA* facilitate bacterial motility inside host cells\textsuperscript{21,22}, with the latter frequently detected in isolates associated with encephalomyelitis in Australia\textsuperscript{19}. These findings validate our analytic approach and ability to detect genetic variations based on geographical origin. The GWAS also identified unappreciated regional variations in well and less well characterised virulence loci (Supplementary table 4), some examples of which are described below.
Filamentous hemagglutinin (fha) is a surface exposed and secreted protein that functions as an adhesin and immunomodulator across different bacterial species. In B. pseudomallei, the number of fha genes varies between isolates, and different combinations of fha genes have been observed between Australia and Thailand. Furthermore, patients infected by B. pseudomallei with a specific fha variant are more likely to have infection associated with positive blood cultures. We identified alternative adhesins/filamentous hemagglutinin variants in the Australasian population (Supplementary table 4). For example, the BURPS668_RS04895 variant in Australasian isolates differed from its non-Australasian ortholog by a group of kmers that clustered in an extended signal peptide for the Type V secretion system, and in hemagglutinin repeat domains (Supplementary Fig. 9a). Such variation may alter protein secretion, binding affinity and specificity.

Intracellular pathogens have evolved various mechanisms for macrophage and immune evasion. Experimental evidence has shown that B. pseudomallei is capable of subverting antigen presentation and macrophage killing via polysaccharide capsule (CPS) and a type III secretion system (T3SS). We identified an Australasian-variant in CPS I (Supplementary Fig 9b), marked by kmers clustered in genes coding for two capsular polysaccharide export ABC transporter transmembrane proteins and putative sulfotransferase. We also identified variation in T3SS between the Australasian and Southeast Asian population (Supplementary Fig 9c). B. pseudomallei carries at least 3 copies of T3SS, but T3SS-3 is considered a virulence factor in mammalian infection. We noted genetic variants in T3SS-3 proteins bsaU, bsaR, bsaP, bsaO, an upstream region of a transcription factor bprR known to activate genes encoding structural components of T3SS-3, and an oxygen-regulated invasion protein orgA in the Australasian population. Infection assays using a macrophage cell line has shown
reduced bacterial escape and lower intracellular bacterial survival of a bsaU mutant\textsuperscript{26}, although the phenotype of geographical variants has not been established.

A distinctive feature of \textit{B. pseudomallei} infection is the formation of multinucleated giant cells (MNGC), which results from cell membrane fusion between infected and uninfected host cells. This enables bacterial cell-to-cell spread while avoiding detection by host immunity. One of the key requirements for MNGC formation is a functional Type 6 secretion system cluster 1 (T6SS-1)\textsuperscript{24}. We detected regional variation that extended from a known Australasian \textit{BmBimA} variant to an upstream region of \textit{virAG} regulator. This locus contains variations in hemolysin-coregulated protein (\textit{hcp}), type VI secretion lysozyme-like protein (\textit{tssE}), and ATP-dependent \textit{clp} protease located on T6SS-1 (Supplementary Fig 9d). It remains to be seen whether region-specific variations in components of T6SS-1 and upstream of the \textit{virA} regulator could affect disease pathogenesis.

In conclusion, our findings link the global dissemination of \textit{B. pseudomallei} to human movement and trade routes. The carrier could have been contaminated soil, water or plants, or humans and other animals with clinical or sub-clinical disease. Given the frequency of \textit{B. pseudomallei} transmission within Asia, it is striking that there appears to have been only one transmission event out of a diverse Australasian population into another geographical location. This might suggest that simple transmission is not sufficient, and that an adaptive bacterial event may also have been necessary. This could reflect the fact that the fauna of Australia and Southeast Asia are significantly different (the Wallace Line\textsuperscript{27}). Our results indicate that movement of people and cargo has led to the dissemination of \textit{B. pseudomallei}, a finding with implications for our increasingly globalised lifestyle. Identification of numerous bacterial genes or gene variants that are geographically segregated provides a rich
resource for biological studies of the basis for region-specific clinical syndromes in melioidosis.
Methods

Bacterial collection and DNA sequencing.
The global *B. pseudomallei* collection sequenced for this study contained 276 isolates from the environment and human disease. The rationale underpinning isolate selection from available global collections was to maximise distribution over time and geography, with representatives from each continent (see Supplementary Fig 2a). A very limited number of isolates had been stored and were available in areas where melioidosis is either uncommon or under-reported based on lack of microbiology infrastructure, which resulted in an unequal geographic representation. DNA libraries were prepared according to the Illumina protocol and sequenced on an Illumina HiSeq2000 with 100-cycle paired-end runs to give a mean coverage of 84 reads per nucleotide (range 35 – 450). Publicly available sequence data for a further 193 isolates (16 reference genomes, 76 Australasian isolates³ and 101 Southeast Asian isolates²) and their accession numbers are also tabulated in Supplementary table 1.

Genome Assembly and Annotation.
To control for potential contamination in each sample with other closely related species, taxonomic identity was assigned to all short reads and assemblies using Kraken²⁸. Multilocus sequence typing (MLST) was derived from Illumina read data by mapping against the MLST sequence archive (http://bpseudomallei.mlst.net/). Unless previously assembled³, *de novo* assembly of short read data was performed using Velvet.³⁰ The kmer size was varied between 60% and 90% of the read length, and the assembly with the best N50 selected. Contigs shorter than the insert size length were filtered out. The sequence data were then used to further improve the assembly. Contigs were iteratively scaffolded using the process described in ³⁰. As a
QC step, reads were mapped back to the assembly using SMALT v. 0.7.4. (http://www.sanger.ac.uk/resources/software/smalt/). The assembly pipeline gave an average total length of 7,139,337 bp (range 6,744,467 – 7,536,799) from 101 contigs (range 72 - 356) with average contig length of 84,361 bp (range 20,098– 192,188 bp) and an N50 of 223,075 (range 37,455 – 1,142,362). Gene predictions and annotations of draft reference genomes as well as other assemblies were performed using Prokka\textsuperscript{31}. On average, 5,980 predicted coding sequences were assigned onto each genome (range 5,701 to 6,671 per each genome), falling within the similar range of a predicted 6,332 coding sequences in the first reference genome K96243 of 7.2 Mb.\textsuperscript{4,32}

\textbf{Pan-genome analysis.}

Based on annotated assemblies, a pan-genome was calculated for all 469 isolates using Roary\textsuperscript{7}. An all-against-all comparison was performed using BLASTP and sequences clustered using a percentage identity of 92%, which was found to be a threshold that optimised specificity and sensitivity in this dataset (Supplementary Fig. 10c and 10d). We identified a total of 25,812 predicted coding sequences, with 4,064 and 21,748 genes assigned to the core (present in 99% of isolates), and accessory (variably present) genome, respectively, which is comparable to that reported previously\textsuperscript{33}. We used rarefaction curves to compare the number of predicted coding sequences as a function of the number of samples detected at different geographies (Fig. 1b and Supplementary Fig. 2c and 2d). A random 1,000 permutations were employed to test our hypotheses. We also tested whether a greater rate of new gene discovery per number of samples sequenced in Australasia was biased by different sampling timeframes or because the sequence data obtained from elsewhere were generated by different sequencing platforms. After sub-sampling the data
Phylogeny based on shared single-copy core genes between *B. pseudomallei* and *B. thailandensis*.

We repeated the pan-genome analysis described above with the inclusion of *Burkholderia thailandensis* genome E264 (accession numbers: NC_007651.1 and NC_007650.1) – a closely related species, which was used as an outgroup to root the tree. This demonstrated that 1,605 single-copy core genes were shared between *B. thailandensis* and *B. pseudomallei*. An approximate maximum likelihood phylogenetic tree was estimated by FastTree version 2.1.3\(^4\) using GTR+CAT (General Time Reversible with per-site rate CATegories) model of approximation for site rate variation and was resampled 1,000 times (Supplementary Fig. 1). The total number of single nucleotide polymorphic sites (SNPs) called was 127,421, of which 69,473 SNPs (54.53\%) marking the differences between *B. thailandensis* and *B. pseudomallei*. This left 57,948 SNPs to resolve the *B. pseudomallei* population structure.

Phylogeny based on core genome mapping of *B. pseudomallei*

A tree was constructed by mapping Illumina sequenced short reads to references using SMALT 0.7.4 (Figure 1a). Fully sequenced chromosomes and long reads sequenced by other platforms\(^3\) were shredded to create 100 bp paired-end reads before mapping. Reads were mapped against the core genome of *B. pseudomallei* strain K96243 (accession numbers BX571965 and BX571966) with bases called and aligned using the method described in \(^35\) and \(^36\). Genetic divergence compared with
the K96243 core genome ranged from 0.73 to 5.61%, and variants were identified at 324,637 SNPs (range 5,650 to 43,221 sites per isolate). A maximum-likelihood phylogeny was estimated with RAxML\textsuperscript{37} using a general time reversible nucleotide substitution model with four gamma categories for rate heterogeneity and 100 bootstrap support.

Hierarchical Bayesian clustering

A tree-independent hierarchical Bayesian clustering with hierBAPS\textsuperscript{38,39} was employed to determine the population structure generated from the core genome mapping alignment. This method allows the population to be sub-divided into groups with closely related genetic backgrounds and allows the recombination detection tool (gubbins) to operate within its best performing range\textsuperscript{40}. Except for the Australasian cluster (Group 1), which contained the highest amount of diversity for each isolate and could not be further sub-clustered, we continued the hierarchical clustering until the diversity observed in secondary or tertiary clusters fell within the limit of recombination detection (Supplementary Fig. 10b). This resulted in 19 groups (Supplementary table 1) for subsequent lineage-specific analyses. Except for Group 15 and a bin cluster (35 isolates), Group 1 - 14 and 16 - 19 each formed a monophyletic group in the phylogeny (Fig 1a).

Analysis of individual lineages.

Evolutionary parameters and date of most recent common ancestors were determined for 19 clusters. For each cluster, closely related reference genomes were chosen for mapping to increase variant calling sensitivity (Supplementary Fig. 3). Where relevant reference genomes were not available as complete chromosomal contigs,
draft reference genomes were created from *de novo* assemblies. One isolate within each of these clusters was selected, assembled and ordered relative to its closest reference using ABACAS v2.5.1\(^{41}\) and ACT\(^{42}\) followed by manual curation. Short reads from all members of each cluster were then mapped against this lineage-specific reference using SMALT 0.7.4. Bases were called and aligned with short insertions and deletions included using the method described in \(^{43}\). Recombination fragments were called and removed from the alignment using Gubbins\(^{11}\). A lineage-specific phylogeny was reconstructed using the remaining variants (Supplementary Fig 4).

**Timeline reconstruction.**

We first tested for a positive correlation between date of isolation and root-to-tip distance obtained from a lineage-specific phylogeny with recombination removed using Path-O-Gen v1.4 (Supplementary Fig. 5). Of 19 clusters, a consistent clock-like behaviour across both chromosomes was observed in a group of American isolates within the African-American cluster and 5 other Asian clusters (groups 4, 5, 6, 7 and 8). Except for group 5 where the number of isolates were too low (n=4) to allow credible estimations, other clusters were analysed by BEAST v1.7\(^{12}\) to determine the clock rate and the time when the most recent common ancestor emerged. We performed model selection on combinations of strict, relaxed log-normal, relaxed exponential, and random clock models and constant, exponential, logistic and skyline population models. For each, three independent chains were run for 50 million iterations, and sampled at every 1,000 generations. Models that failed to converge based on visual inspection\(^{44}\) of the trace files or had effective sampling size (ESS) values < 200 for key parameters were discarded. Stepping-stone and path-sampling analyses did not show appreciable differences between clock models, potentially
suggesting that there may be insufficient rate variation within each group to warrant
the use of a complex clock model. Thus, the strict clock with fewest parameters was
employed to avoid over-fitting of parameters as suggested in\textsuperscript{45}. We used the Bayesian
skyline model as the tree prior to describing demographic history. Except for
chromosome I of group 4 which did not achieve a credible ESS, the time calibrated
phylogenetic trees, clock rates and time since most recent common ancestor
(TMRCA) of estimated clusters are reported in Supplementary Fig. 6.

Due to a small sample size used for each estimated cluster (American isolates within
group 19 = 9 isolates, group 4 = 11 isolates, group 6 = 24 isolates, group 7 = 9
isolates, and group 8 = 6 isolates), we also performed a date-randomised test as
described in\textsuperscript{46} to estimate the rigour of the true temporal signals compared to noise.
For each tested cluster, we performed 1,000 permutations with the true date, but
randomised root-to-tip distance. Regression coefficient $R^2$ of the true data was ranked
and compared to $R^2$ of the randomised data (Supplementary Fig. 5). Ranks of the true
signals ranged from 34\textsuperscript{th} (group 6 chromosome II) to 97\textsuperscript{th} (group 8 chromosome II),
suggesting that noises had an effect on a small dataset. Aside from small sample size,
our clock rate on each chromosome of the clusters estimated by BEAST is consistent
with previous estimate in \textit{Burkholderia} species\textsuperscript{47} and other bacteria\textsuperscript{35,48-50}. This
suggests that the results generated here are non-random.

\textbf{Ancestral state reconstruction on geographic locations of Southeast Asian
isolates.}

Ancestral reconstruction was performed on the maximum likelihood global core
genome phylogeny to assess the connectivity of isolates, and infer which population
might act as source versus sink in Southeast Asia. To avoid sampling bias, we sub-sampled the phylogeny so that there were equal numbers of isolates from Thailand, Laos, Cambodia, Vietnam, Malaysia and Singapore (n=15 for each countries), and permuted 1,000 times. Countries contain less than 15 isolates were excluded. We treated countries as discrete geographic characters. For each sub-sampled tree, we used stochastic character mapping *make.simmap* available in R package phytools to estimate both the transitions between different geographical characters and the total time spent in each geographical character. Stochastic mapping was performed under an asymmetric model of character change for 1,000 simulations.

To assess the connectivity of isolates, we categorised geographical characters into two groups based on geographical proximity. The Mekong sub-region represents countries bordered by the Mekong river including Thailand, Laos, Cambodia and Vietnam; the Malay sub-region comprises Malaysia and Singapore. Changes between geographical characters were counted after grouping into two categories: 1) transitions within the same sub-region, and 2) transitions between sub-regions. The occurrence of transitions within and between the two sub-regions was compared using a two-tailed Mann-Whitney U test (Supplementary Fig. 7b). To infer which population might act as the source, we compared the time spent in the Mekong and Malay sub-regions and compared this using a two-tailed Mann-Whitney U test (Supplementary Fig. 7c). The choice of non-parametric Mann-Whitney U test over parametric test was due the violation of normally distributed data.

**Identification of distinct genes/variants in Australasian and Southeast Asian populations.**
Kmer-based GWAS without correction for population structure.

We first considered the optimal approach to perform a GWAS for *B. pseudomallei*. Given the high level of genomic plasticity and large accessory genomes (Fig. 1c), we concluded that a GWAS based on core genome SNPs as used elsewhere would be sub-optimal as this fails to capture the extent of genetic variation. Instead, we used kmers (DNA words of length k) as an alternative to a SNP-based analysis. Unlike a traditional GWAS where genetic causes of particular phenotypes were identified while adjusting for population stratification, we employed GWAS to search for genetic markers in the Australasian and Southeast Asian populations, some of which may intrinsically define population structure. A control for population structure was thus omitted. Two independent GWAS runs were performed to search for variable kmers in the Australasia population alone (Australasia GWAS), and the Southeast Asian population alone (SEA GWAS). For both GWAS runs, the data was randomly divided into a discovery and a validation dataset. The Australasia GWAS comprised a set of 80 Australasia and 200 non-Australasian isolates, and was validated in 57 Australasian and 132 non-Australasian isolates. Similarly, the SEA GWAS comprised a random set of 180 Southeast Asian and 105 non-Southeast Asian isolates, and was confirmed using 114 Southeast Asian and 65 non-Southeast Asian isolates. We used the reference-independent GWAS pipeline Seer by Lees et al. to search for kmers with region-specific patterns. All kmers of length 9-100 bp were scanned from all assembled reads using fsm-lite (https://github.com/nvalimak/fsm-lite). Only kmers seen at 5-95% of the total population were retained to reduce false positives from testing underpowered kmers. Seer was performed on the discovery data using geographical origins of isolates (Australasia/ non-Australasia or Southeast Asia/ non-
Southeast Asia) as binary phenotype \( (Y) \) and the presence/absence of each kmer as tested genotype \( X \):

\[
\log \left( \frac{Y}{I-Y} \right) = X\beta
\]

The direction of association (positive or negative) is described by \( \beta \). Kmers with a conservative cut-off p-value < \( 10^{-8} \) in the logistic regression were considered further as suggested in \(^{18}\). Australasia and SEA GWAS respectively yielded 77,787 and 43,663 kmers that were positively or negatively associated with Australasia or Southeast Asia populations (Supplementary table 2). Among these, 42,521 kmers that were positively associated with Australasia were negatively associated with Southeast Asia (Supplementary Fig. 11). Kmers that reached significance in the discovery data were confirmed the discovery data. To aid visualisation, the frequencies of 5,000 randomly chosen kmers from the Australasia and SEA GWAS in the validation data have been plotted in Supplementary Fig. 11.

**Mapping and kmer clustering.**

Significant kmers were searched for an exact match in \textit{de novo} assemblies and fully sequenced chromosomes using BLAT v. 34 \(^{55}\) with minimum match and score adjusted to cater for low complexity kmers as below.

\texttt{blat assembly kmers.query –minMatch=1 –minScore=10 output}

To facilitate biological interpretation, kmers were grouped into clusters based on their genetic distance. We defined the size of operons based on the length of transcription fragments reported in \(^{56}\). Any kmers located within 7.68 kb (the size of an operon covering 95\textsuperscript{th} percentile of transcription fragments) were grouped together into a locus. On average, each locus had a median of 66 kmers (range 2 -11,072 kmers), with the size of the loci ranging from 40 – 70,684 bp (Fig 3a). The binary patterns in
size of regional-specific loci (Fig 3a, top histogram) likely reflect different scales of variation, with smaller and larger peaks corresponding to small-scale differences (including SNPs) and large-scale differences (including regions of mobile genetic elements incorporated via homologous recombination or site specific recombination), respectively. As the GWAS was not corrected for population structure, we further tested whether the predicted loci were subjected to clonality. The presence and absence of each locus (measured by % of detected kmers) were plotted against the phylogeny. Their scattering patterns across multiple branches suggested that region-specific loci were not strictly driven by a clonal population structure (Supplementary Fig. 8).

**COG, GO and pathway terms found in region-specific loci.**

We annotated the biological properties of kmers within coding regions using information from the functional categories (COG term), Gene Ontology (GO term), and pathway data (KEGG, InterPro and UniPathway), available from 57. The reference genome Bp668 contained 40,986 out of 78,929 region-specific kmers, of which 23,565 overlapped with coding regions. One-sided Fisher’s exact test was used to search for COG, GO and pathway terms in kmers that showed significant departure from random expectation in the Bp668 genome. We tested kmers enrichment in 22 COG terms, 1,485 GO terms, and 408 pathway terms using a strict Bonferroni correction with a required p-value of 0.01/1,915 = 5.22x10^{-6}. Significant COG terms were highlighted in Fig. 3b. Additional GO and pathway enrichment analyses are discussed in the supplementary note. Terms with significant deviation are tabulated in Supplementary table 5.
Statistics and visualisation.

Visualisation of phylogenetic trees and statistical analyses were performed in R\textsuperscript{58}, iTol\textsuperscript{59}, and FigTree v 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Correspondence and request for materials should be addressed to Claire Chewapreecha (echewapreecha@gmail.com)

Data availability.

New sequence data for the study isolates have been deposited in the ENA under study accession number ERP001193 and ERP002658, with the accession numbers for individual isolates listed in Supplementary table 1.

References

Note: number 1 - 27 are in text references


Holden, M. T. et al. Genomic plasticity of the causative agent of melioidosis, 
*Burkholderia pseudomallei*. *Proc Natl Acad Sci U S A* 101, 14240-14245, 


Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and postanalysis of large phylogenies. *Bioinformatics* 30, 1312-1313,


Ooi, W. F. et al. The condition-dependent transcriptional landscape of


Acknowledgements

The authors thank the Wellcome Trust Sanger Institute library construction, sequence and core informatics teams, and Elizabeth Blane for their technical support. We thank Drs Tanistha Nandi and Patrick Tan at Genome Institute of Singapore; and Drs Erin Price, Derek Sarovich at Menzie School of Health Research, Australia for providing
access to publically available WGS data. We thank the following people who provided isolates or DNA: Professor Nicholas Day, MORU, Faculty of Tropical Medicine, Mahidol University; Drs Paul Newton, Manivanh Vongsouvath, Mayfong Mayway, Viengmon Davong, Olay Lattana, Catrin Moore, Sayaphet Rattanavong and the directors and staff of Mahosot Hospital, Vientiane, Lao PDR; Dr Varun Kumar, Ankor Hospital for Children, Siem Reap, Cambodia; Dr James Campbell, Oxford University Clinical Research unit, Ho Chi Minh City, Vietnam; Dr Hui Suk Wai, Ocean Park Corporation, Hong Kong SAR, China; Mr Chun Kham and Dr Thong Phe, Sihanouk Hospital Centre of Hope, Phnom Penh, Cambodia; Dr Joost W. Wiersinga, Academic Medical Center (AMC), Amsterdam, the Netherlands; Professor Jan Jacobs, ITM, Antwerp, Belgium; Dr Julie E. Russell, National Collection of Type Cultures, UK; Dr Ty Pitt, NHS Blood and Transplant, UK; Dr Daniel Godoy, Imperial College, UK; Dr Stephane Emoret, Geneva University Hospitals, Switzerland; Dr Susan Morphet, Middlemore Hospital, New Zealand; and Dr Jay Gee, CDC, USA. C.C. is a Sir Henry Wellcome post-doctoral Fellow (grant ref: 107376/Z/15/Z). J.C., M.V., Z.Y. were supported by the COIN Centre of Excellence and Z.Y. by a HIIT post-doctoral fellowship. A.E.M. is supported by Biotechnology and Biological Sciences Research Council grant BB/M014088/1. B.G.S. was supported by the Wellcome Trust grant WT089472. This publication presents independent research supported by the Health Innovation Challenge Fund (WT098600, HICF-T5-342), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health or Wellcome Trust. This project was also funded by a grant awarded to the Wellcome Trust Sanger Institute (098051).
**Author contributions**

A.T., B.D.S., S.L.H., C.B., M.M., V.W., D.L., R.P., B.G.S., P.K., D.A.B.D. and B.J.C. collected and provided the samples for the study. C.C. designed and performed the analyses. M.T.G.H, S.R.H., A.E.M., J.C., J.P. and G.D. designed and contributed materials and analysis tools. M.V., N.V., Z.Y., and J.C. performed kmers based analyses in the first draft. C.C. performed kmers based analysis in the revised draft. Z.Y. and J.C. performed cluster analyses. S.J.P. was responsible for management of the study. S.J.P. and C.C. wrote the paper with input from all authors. All authors approved the manuscript prior to submission.

**Competing financial interests**

The authors declare no competing financial interests.
Figure legends

Fig. 1 The phylogeny and pan-genome of *B. pseudomallei*

Differences in level of bacterial diversity across different geographical origins: Australasia (green), Asia (yellow, cyan, and magenta for (a) and yellow for (b and c)), Africa (blue), America (red), and Europe (star). (a) A core SNP-based maximum likelihood phylogeny of 469 genomes with geographical origins highlighted. The tree was rooted on *B. pseudomallei* MSHR5619 - the most genetically distant isolate based on pairwise SNP distance (see methods and Supplementary Fig. 9). The outer ring represents population clusters based on BAPS hierarchical clustering (Group 1 – 19). Apart from Group 15, which is paraphyletic and marked by two black arrows, other groups each form a monophyletic branch. (b) Pan-genome accumulation curve representing rates of new gene discovery in isolates collected from different geographical origins. (c) Summary of core and accessory genomes of isolates grouped by geographical origins.

Fig. 2 Timeline of trans-continental and sub-regional spread of *B. pseudomallei*

(a) Estimated time when the most recent common ancestor (MRCA) of each cluster emerged. Time (black dots) and 95% highest posterior density (horizontal line) were estimated by BEAST for those clusters with temporal signals. Estimations were performed separately for chromosome I (solid lines), and II (dotted lines). Overlapping estimations between the two chromosomes provide further confidence in the time interval in which the MRCA emerged. The estimation for chromosome I of group 4 did not reach a credible effective sample size and was excluded. (b) Transatlantic slave trade routes and sampling locations of African and American isolates. Each dot represents the geographical origin of isolates used for the time
estimation with the size proportional to the number of isolates. (c) The geographical landscape and isolates used to determine sub-regional connectivity. Isolates representing six Southeast Asian countries were plotted on the map, highlighting the geographical proximity of the Mekong group, and the Malay group.

Fig. 3 Region-specific genetic signatures
(a) Size and functional categories of genes located in region-specific loci. The horizontal axis shows the size of loci (bp), with a frequency histogram plotted on the top. The histogram peaks at 100-300 bp (top-left) and 10 kb (top-right), potentially highlighting small-scale (single polymorphic differences) and large-scale (likely introduced via horizontal gene transfer) variation, respectively. The vertical axis displays the number of kmers used to define each locus, with a histogram plotted on the right. Circles represent each locus, the size of which is proportional to the number of genes identified in each locus. Circle colour represents the functional category (COG), with the colour scheme shown in (b). Colour was selected based on the most prevalent COG assigned to genes in a particular locus. Black circles are loci within intergenic regions. (b) Functional categories of genes (COG) localised in region-specific loci. Asterisks highlight terms with heightened frequency compared to random expectation from a reference genome (see methods). *denotes terms with p-value <10^{-9}, while ** denotes terms with p-value <2.2x10^{-16}.

Supplementary Fig. 1 Phylogeny based on shared single-copy orthologous genes between Burkholderia pseudomallei and Burkholderia thailandensis
An approximate maximum likelihood phylogenetic tree was constructed using 1,605 single-copy orthologous genes conserved between B. pseudomallei and B.
The tree was rooted on *B. thailandensis*. *B. pseudomallei* isolates are coloured by geographical origins for each isolate using the colour scheme described in Figure 1a.

**Supplementary Fig. 2** Geographically distinct rates of new gene discovery are not affected by different sampling timeframe or sequencing platforms. Rates of new gene discovery were checked against distributions of sample collection by year and assembly quality. Rarefaction curves showed no change in the curve trajectory, regardless of potential bias in sampling timeframe or different sequencing technologies. (a) Histograms summarising sample collection by year. Horizontal dotted lines represent the threshold at which the number of sub-sampling isolates falls below 20, 40 and 60 isolates per decade, respectively. (b) Histograms summarising the number of contigs obtained for isolates, which are either Illumina-sequenced, PacBio-sequenced or high quality reference genomes (Supplementary table 1). Vertical dotted lines highlight the threshold where the number of contigs fall below 10, 100 and 200 contigs. (c) Rarefaction curves generated by sub-sampling the data using thresholds marked in (a). (d) Rarefaction curves generated by sub-sampling the data using thresholds marked in (b). The same colour scheme for Fig. 1 was used to highlight the geographic source of each isolate.

**Supplementary Fig. 3** Lineage-specific analyses enhanced by mapping against closely related references. Where reference genomes were members of BAPS clusters (a, f, i, m, n, p, q, r), these were used as mapping references. Where reference genomes were not available, new references were generated from assemblies and ordered to their close reference genomes (b, c, d, e, g, h, j, k, l, o, s). Sequence
similarity between ordered contigs and their close reference are shown with different shades of red (dark red = 100% identity). Contig breaks (vertical lines) in new references were marked as annotated in (b). Grey rectangles beneath previously available or newly generated references represent genomic islands identified in Tuanyok et al. 2008. Recombination fragments identified by Gubbins were plotted (black dots) on reference chromosomes, allowing a comparison of recombination landscape between chromosomes I and II.

Supplementary Fig. 4 Recombination-free phylogeny of each lineage divided by chromosome. (a - o) Final phylogenies for chromosomes I and II with recombination removed for Groups 1 - 3, 5 & 9 - 19. Final phylogenies for Groups 4, 6, 7, 8 and American isolates within Group 19 are shown in Supplementary Fig. 5 as time-calibrated phylogenies.

Supplementary Fig. 5 Testing for molecular clock signal using Path-O-Gen.
(a) Correlation between time and root-to-tip divergence for all 469 genomes shown in Fig. 1a. The regression coefficient ($R^2$) was used to estimate the fit of the data to a strict molecular clock. The plot rejects the influence of sampling time over the amount of root-to-tip diversity in the whole dataset. (b) Correlations between time and root-to-tip divergence in individual groups and separate chromosomal compartments. Phylogenies with recombination removed (Supplementary Fig. 3) were used to determine root-to-tip divergence. For all separate panels, a timeline (in years) and root-to-tip divergence are represented on the horizontal and vertical axes, respectively. Groups with potential molecular clock signals across both chromosomes are highlighted in yellow. For each of these groups, a percentile rank of $R^2$ from the
true signal against 1,000 randomised signals is documented underneath its $R^2$. Group 5 shows a clock signal but only contains 4 isolates and cannot be reliably assessed.

**Supplementary Fig. 6 BEAST estimation for lineages with clock signals.**

(a – e) BEAST trees for groups 4, 6, 7, 8, and American isolates within group 19, with error bars showing 95% HPD. Trees constructed from chromosome I and II were positioned on the top and bottom of each panel, respectively. The estimation for group 4 chromosome I did not reach a credible ESS and was excluded from the analysis. (f) A table summarising rates of nucleotide substitution and time since most recent common ancestor (TMRCA) estimated by BEAST. The rates of nucleotide substitution are consistent with the previous estimate in *Burkholderia* species described in Lieberman *et al.* 2011.

**Supplementary Fig. 7 Patterns of sub-regional spread.** (a) A global phylogeny with Southeast Asian isolates highlighted. Countries bordered by Mekong river (Cambodia, Laos, Thailand and Vietnam) were collectively termed the “Mekong” sub-region, while countries located on the Malay peninsula (Malaysia and Singapore) were termed the “Malay” sub-region. (b) Boxplots summarise the proportion of node transitions observed between the same and across sub-regions: from Mekong to Mekong, from Malay to Malay, from Malay to Mekong, and from Mekong to Malay, respectively. Each black dot represents the result of 1,000 sub-samplings of the phylogeny in (a), with 1,000 permutations of stochastic character mapping applied to each sub-sampled phylogeny. The boxplots represent lower quartile, median, and upper quartile respectively (c) Boxplots summarise the proportion of time spent in
each state using the same sub-sampling and stochastic character mapping analyses described in (b).

Supplementary Fig. 8 Region-specific loci are not solely driven by population clonal structure. A heat-map of region-specific loci observed in Australasian (n=468) or Southeast Asian (n=14) populations was plotted against the phylogeny in Fig 1a (far left). The second and third columns from the left summarise the geographical origin of each isolate, and whether used in the GWAS discovery or validation dataset. For each locus, the presence or absence was defined by the total number of detected kmers over the number of kmer used to construct each locus (on a scale of 0 to 1). The presence, or score 1 is shown in green and yellow for Australasia- and Southeast Asia specific loci respectively. The absence, or score 0 of these loci is represented in grey. Reference genomes and isolates assembled into two full chromosomes are denoted ** and * respectively. Patterns of region-specific loci observed in these genomes are unlikely influenced by assembly errors.

Supplementary Fig. 9 Examples of region-specific kmers enriched in different B. pseudomallei populations. Kmers associated with: (a) filamentous hemagglutinin; (b) capsular polysaccharide biosynthesis locus (CPS-1); (c) Type 3 secretion system cluster 3 (T3SS-3); and (d) Type 6 secretion system cluster 1 (T6SS-1) and *Burkholderia* intracellular motility *bimA*. For each figure, the horizontal axis is a genome coordinate with predicted genes plotted in the bottom of the diagram. The vertical axis is the frequency of region-specific kmers detected, with green, yellow and grey circles representing their frequency in Australasian, Southeast Asian and other populations, respectively. These examples confirm enrichment of kmers in the
Australasian population but not in the Southeast Asian population. Genes are annotated in black. Protein domains are annotated in grey with shaded grey bars. A well-characterised *Burkholderia mallei*-like *bimA* variant predominantly detected in Australasia is highlighted in pale green in (d).

**Supplementary Fig. 10** Determination of levels of hierarchical BAPS clustering, and cut-off identity for pan-genome analysis. Hierarchical clustering was performed to achieve final level clusters with the number of SNPs falling within the recombination detection threshold, as described in Croucher *et al.* 2013\(^{40}\). (a) Distribution of pairwise SNP distance between isolates in the entire collection (white), and isolates categorised within the same Bayesian clustering (primary-red, secondary-blue, tertiary-magenta). Based on this plot, isolate MSHR5619 was determined as the most distant isolate (black triangle) and used to root the tree in Fig. 1a. (b) Final number of SNPs per branch resulting from primary, secondary and tertiary clusters used in the analyses. Dotted lines shown in (a) and (b) represent the accuracy of recombination detection by Gubbins. The cut-off for nucleotide identity used to determine orthologous groups in the pan-genome analysis was selected to maximise stringency and clustering information. We tried several cut-off thresholds ranging from 60% to 98% identity and compared a strict core pan-genome (100% present in all genomes) against the reference genome K96243. (c) The total number of base pairs (bp) covered by a core pan-genome on K96243 chromosome I, II and combined chromosomes. (d) The total number of predicted coding sequences (CDS) covered by a strict core pan-genome on K96243 as in (c). A threshold of 92% cut-off (red triangles) was selected.
Supplementary Fig. 11 Region-specific kmers discovered from Australasian and Southeast Asian GWAS analyses.

(a) Venn diagram summarises the number of significant kmers detected in the discovery data of Australasia-specific and Southeast Asia-specific GWAS. Kmers positively associated with the Australasian alone, positively associated with the Southeast Asian alone, or positively associated with the Australasian population but negatively associated with the Southeast Asian population are coloured in green, yellow and pale green, respectively. (b-d) Frequency of kmers detected in the validation dataset. Each set of 5,000 kmers was randomly chosen from the Australasia GWAS and Southeast Asia GWAS to allow visualisation. Horizontal axes represent kmer frequency in Australasian (b & c), and Southeast Asian (d & e) populations, and kmer frequency in control population (non-Australasia in b & c, and non-Southeast Asia in d & e) are presented on the vertical axes. Each dot represents a kmer with the colour coded for an association score (b & d), and direction of association (c & e). The association score is based on a negative log of Bonferroni adjusted p-value. Direction of association is dictated by the beta value where negative and positive scores represent negative and positive associations, respectively.

Supplementary table 1 Epidemiological data. Isolate and accession codes associated with data deposited in the European Nucleotide Archive (ENA) for newly sequenced isolates, or NCBI for public data used in this study.

Supplementary table 2 Frequency, association scores and direction of association for significant kmers identified in Australasian and Southeast Asian GWAS analyses. A table summarises region-specific loci and their kmer members with kmer
frequency in the Australasian, the Southeast Asian, and other populations. A 
Bonferroni adjusted p-value and beta score from a regression performed for each 
kmer is also reported.

Supplementary table 3 Raw kmer sequences.
A total of 78,929 raw kmer sequences from the Australasia and Southeast Asia 
GWAS. Kmer IDs are the same as Supplementary table 2.

Supplementary table 4 Coding sequences found in region-specific loci.
Summary data of all annotated coding sequences (CDS) located in 468 Australasia- 
and 14 Southeast Asia-specific loci. Exact kmer positions and kmer-containing CDS 
can be found by mapping raw kmer sequence in Supplementary table 3 onto each 
region-specific locus.

Supplementary table 5 COG, GO and pathway terms with significant kmer 
enrichment.
Summary data for kmer enrichment analyses on functional categories of genes 
(COG), gene ontology (GO), and pathway terms. Terms reaching statistical 
significance following Bonferroni correction are highlighted in green.
addition of new genomes
new CDS discovered

Fig 1

a  0.01

b

c

Number of coding sequences observed

0 10000 20000

Total
Australasia
Asia
Africa
America

Core  Accessory

Australasia
Southeast Asia
East Asia
South Asia
Africa
America
Europe
American isolates within group 19 (Africa-America)

Group 4 (Singapore-Malaysia)
Group 6 (Thailand-Laos)
Group 7 (Malaysia-Singapore)
Group 8 (Malaysia-Singapore)

Year:

1650 1700 1750 1800 1850 1900 1950 2000

b

Slavery routes in 1650–1850
African isolates
American isolates