Genetic diversity, determinants, and dissemination of *Burkholderia pseudomallei* lineages implicated in melioidosis in northeast Thailand

Rathanin Seng[1], Sarunporn Tandhavanant[1], Natnaree Saiprom[1], Rungnapa Phunpang[1], Chalita Chomkatekaew[2], Janjira Thaipadungpanit[3], Elizabeth M Batty[2,4], Nicholas R Thomson[5], Wasun Chantratita[6], T. Eoin West[1,7,8], Claire Chewapreecha[2,3,5]*, Narisara Chantratita[1,2]*

[1] Department of Microbiology and Immunology, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.

[2] Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.

[3] Department of Clinical Tropical Medicine, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.

[4] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK.

[5] Parasites and Microbes, Wellcome Sanger Institute, Cambridge, UK.

[6]Center for Medical Genomics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[7]Division of Pulmonary, Critical Care & Sleep Medicine, Department of Medicine, University of Washington, Seattle, Washington, USA.

[8] Department of Global Health, University of Washington, Seattle, Washington, USA.

Co-corresponding authors: Narisara Chantratita (narisara@tropmedres.ac, +66(0)819099772) and Claire Chewapreecha (claire@tropmedres.ac, +66(0)952743357)

Abstract

Melioidosis is an often-fatal neglected tropical disease caused by an environmental bacterium *Burkholderia pseudomallei*. However, our understanding of disease-causing bacterial lineages, dissemination, and their adaptive mechanisms remains limited. To address this, we conducted comprehensive genomic analysis of 1,265 *B. pseudomallei* isolates collected from nine hospitals in northeast Thailand between 2015 and 2018, presenting the most densely sampled collection to date. Our study identified four dominant lineages with unique gene sets enhancing bacterial fitness, indicating lineage-specific co-colonisation strategies. Crucially, recombination was found to drive lineage-specific gene flow. Transcriptome analysis of a representative isolate from one lineage revealed upregulation of lineage-specific genes under nutrient-depletion. Our study also demonstrated the impact of river systems, monsoons, and typhoons in shaping the geographical dispersal of *B. pseudomallei*. Collectively, our findings underscore the importance of bacterial persistence in the environment as a prelude to exposure and infection, thereby providing insights for melioidosis prevention and control strategies.

Introduction

Melioidosis, a severe infectious disease, is estimated to affect 165,000 cases globally per year, of which 89,000 is fatal[1]. The disease is caused by an environmental Gram-negative bacillus *Burkholderia pseudomallei*, which can be found in soil and contaminated water across tropical and sub-tropical regions. The disease has been historically underreported due to limited availability of microbiology laboratory infrastructure required for culture-confirmed melioidosis diagnosis, particularly in lower- and middle-income countries. Although recent reports of increasing melioidosis cases in South Asia, Southeast Asia, East Asia[2-7] and Australia[8,9] may partly be attributed to improved microbiology infrastructure and growing disease awareness, the cases have often been associated with severe weather events[10-12]. While climate likely influences human encounters with *B. pseudomallei*, further investigation is needed to fully understand the mechanisms linking environmental factors to melioidosis epidemiology.

Understanding the population structure, dissemination and adaptation of *B. pseudomallei* in these climatically challenged endemic regions requires a large-scale, chronologically, and densely sampled genetic dataset. Previous studies, albeit limited in sample size, have demonstrated that *B. pseudomallei* dissemination is driven by both anthropogenic and environmental factors[13-16]. Water streams[17,18], monsoons, typhoons and cyclones[10,12,19,20] were identified as significant contributors to bacterial dissemination, highlighting the importance of bacterial persistence across a range of environmental conditions. *B. pseudomallei* exhibits

remarkable survival capabilities across diverse environments, ranging from nutrient-rich to nutrient-depleted soil and water[21-26] thereby enabling the bacterium to thrive in various ecological niches. Previous studies also noted the temporal and geographical co-existence of multiple *B. pseudomallei* lineages[13,14,27]. However, little is known about their distinct genetic content and adaptive strategies. An identification of lineage-specific genes and genetic variants associated with bacterial persistence and disease escalation will be essential to develop the disease control strategies.

In this study, we performed a population genomics analysis of 1,265 *B. pseudomallei* isolates from melioidosis patients in northeast Thailand. We elucidated the population structure, investigated the dissemination patterns, genetic diversity, and identified genetic determinants associated with dominant lineages. Additionally, we explored the biological functions and expression conditions of these genetic determinants in a representing strain from one dominant lineage. This provides insights into the strategies employed by *B. pseudomallei* lineage to ensure their successful persistence in the environment, ultimately leading to exposure and infection.

Results

Population structure analysis revealed the successful *B. pseudomallei* lineages

To define the population structure of *B. pseudomallei* in hyperendemic area, we employed two independent approaches. We constructed a maximum-likelihood (ML) phylogeny and performed PopPUNK analysis of 1,265 genomes of *B. pseudomallei* obtained from patients with melioidosis residing in nine provinces in northeast Thailand during July 2015 to December 2018 (**Fig. 1a**). Both PopPUNK and phylogenetic analysis gave consistent results with these genomes assigned into 101 closely related lineages. By size, we identified four dominant lineages (lineage 1 to 4) comprising more than 100 isolates each, accounting for 54.7% of the samples. Additionally, there were 19 intermediate lineages (lineage 5 to 23) with at least 10 isolates per lineage, accounting for 33.3% of the samples. The remaining 78 lineages (lineages 24-101) were classified as minor, representing 11.8% of the samples (**Fig. 1b and 2; Supplementary data 1**). To validate our classification, we compared the pairwise distances of core single nucleotide polymorphism (SNP) between isolates within and between lineages. The results showed that the SNP distance was lower within lineages compared to between different lineages (median, 458 versus 1,124 SNPs; collective p-value $< 2.2 \times 10^{-16}$), confirming the relationship defined by PopPUNK and the phylogeny (**Fig. 1c**). Furthermore, our population structure also matched previously established multi-locus sequence typing (MLST) classification.

Among the 1,220 isolates that could be assigned to known sequence types (STs), we observed dominant STs that correlated with dominant lineages (**Fig. 2; Supplementary data 1**). For example, lineage 1 was associated with ST 70, 220 and 512, lineage 2 with ST 17, 23, 206, 306, 307, 520 and 526, lineage 3 with ST 9, 10 and 538, and lineage 4 with ST 46, 48, 56 and 167 (**Fig. 2**). Throughout the sampling period, the four dominant lineages consistently dominated in terms of size and persistence, indicating their robust and stable presence in the sampled population (**Fig. 1d**). To understand the bacterial dissemination route and lineage-specific adaptive strategy, our analysis was next focused on the dominant lineage 1 (n = 267), lineage 2 (n = 183), lineage 3 (n = 129), and lineage 4 (n = 114).

Genetic evidence highlighted patterns of *B. pseudomallei* dissemination in Northeast Thailand

The dominant lineages of *B. pseudomallei* were found to be dispersed across the nine studied provinces (**Fig 3a**). To delve deeper into the dissemination patterns within each of these lineages, we constructed lineage-specific phylogenies, reconstructed ancestral histories of provincial origins, and quantified the number of inter-provincial transmissions (**Supplementary Fig. 1**, see methods). In most province-to-province pairs (13 out of 21 inter-provincial transmissions), we observed bidirectional transmission with similar number of transmission events but acting in opposite directions (**Fig. 3b**). This suggests the absence of geographical barrier that preventing the spread of *B. pseudomallei* in northeast Thailand. However, a subset of inter-provincial transmissions (8 out of 21 pairs) displayed unidirectional patterns across lineages including the “north-to-south”, “northeast-to-southwest”, and “east-to-west”. The “north-to-south” dissemination was supported by a significantly higher number of transmissions from “Nakhon Phanom-to-Mukdahan”, “RoiEt-to-Surin”, “Udon Thani-to-Buriram”, and “Udon Thai-to-Surin” compared to the reverse direction (each lineage transmission Mann Whitney U Test p-value $< 2.2 \times 10^{-16}$, **Fig. 3b**). This pattern aligns with the “north-to-south” flow of the great Mekong River and its inland river system, which transverses multiple countries in Southeast Asia. The "northeast-to-southwest” transmission was evident through an elevated number of transmissions from “Roi Et-to-Buriram”, “Mukdahan-to-Surin” and “Mukdahan-to-Buriram” (each lineage transmission Mann Whitney U Test p-value $< 2.2 \times 10^{-16}$). This dissemination pattern may be influenced by the regular north-east monsoon, which typically travels from northeast to southwest of Thailand during the dry season may in part contribute to this pattern of dissemination. Additionally, the “east-to-west” dissemination, exemplified by a significantly

greater transmission from "Surin-to-Buriram" than compared to the reverse direction (each lineage transmission Mann Whitney U Test p-value < 2.2 x 10^{-16}). This pattern be influenced by typhoons that occasionally occur in the regions, blowing in an anti-clockwise direction. It is highly likely that multiple factors including the network of inland rivers and canal systems[28], the regular monsoons[29], and increasingly frequent occurrence of typhoons[30] had collectively shaped the dissemination pattern of *B. pseudomallei* in northeast Thailand (**Fig. 3c**). Anthropogenic activities, such as human migration between the provinces, may also contribute to the observed pattern; however, we did not have access to human movement data to investigate this aspect thoroughly.

Genetic markers potentially contributing to the emergence of successful *B. pseudomallei* lineages

The co-existence of four dominant lineages in a spatiotemporal context suggests the presence of multiple adaptive strategies. We next sought to identify genes and SNPs that were present in all, or in each of the isolates that form four dominant lineages (n = 693 isolates) but absent in others (n = 572 isolates) (see methods). Dominant lineage-specific genes were defined as accessory genes present in ≥95% of isolates within any of four dominant lineages, but present in ≤5% (strict threshold), ≤10% (intermediate threshold), or ≤15% (relaxed threshold) of isolates outside these lineages. Out of the 7,220 accessory genes identified in this population (see methods), 17, 41, and 68 dominant lineage-specific accessory genes were identified based on the strict, intermediate, and relaxed threshold, respectively. These genes exhibit lineage-specific patterns, with many being specific to either lineage 1 or lineage 3, while several were shared between lineage 2 and 3 (**Supplementary data 2; Fig. 4a; Supplementary fig. 2-3a).**

We also investigated SNPs present in all, or each of the dominant lineages. Core SNPs were reconstructed onto the phylogeny shown in **Fig. 2** using the maximum parsimonious method. Of 136,477 core SNPs used to define the entire population; we identified 654 dominant-lineages specific SNPs spanning 381 core genes. The majority of these dominant lineage-specific SNPs were found exclusively in lineage 1 (54 core SNPs in 34 genes), lineage 3 (587 core SNPs in 342 genes) or shared between lineage 1 and 3 (13 core SNPs in 10 genes) (**Supplementary data 2-3; Supplementary fig. 3b & c**). Thus, the dominant lineages of *B. pseudomallei* in northeast Thailand may be characterised by the presence of 68 accessory genes (using the relaxed threshold) or the presence of SNPs in 381 core genes (**Supplementary fig. 3d**).

To gain insights into the functions of these dominant lineage-specific genes, we annotated them using clusters of orthologous groups (COG) terms. COG assignment could be made to 18 (26.4%) of the 68 accessory genes and 292 (76.6%) of the 381 genes carrying dominant lineage-specific SNPs. However, our analysis did not reveal any significant enrichment of genes associated with specific COG terms (**Fig. 4c**; **Supplementary Data 4**).

Lineage-specific genes were selectively expressed

To gain a deeper understanding of the functionality of lineage-specific genes, including both accessory genes and genes carrying lineage-specific SNPs, we investigated their expression patterns across various environmental and infection conditions. These conditions encompassed growth, anaerobic condition, desiccation and rehydration, supplementation with amino acids, divalent cations, iron and human factors, temperature, osmotic and oxidative stress, UV irradiation, nutrient deprivation, antibiotic treatment, changes in pH, exposure to heavy metals and both *in vitro* and *in vivo* infections. We utilised the condition-wide transcriptome data[31,32] available for a reference strain in lineage 3, named "K96243" (a total of 65 conditions, **Fig 5).** This enables us to compare the expression profile of the lineage-specific genes with other genes in the genome. Our analysis revealed that both the accessory genes specific to lineage 3 (n = 29 genes) and the genes carrying SNPs specific to lineage 3 (n = 345 genes) were not expressed as constitutively as the remaining genes in the K96243 genome (n = 5,840 genes). The median number of conditions in which the accessory genes specific to lineage 3, genes carrying SNPs specific to lineage 3, and the remaining genes expressed were 15, 23, and 31 out of 65 conditions, respectively (**Fig 5a**). This observation was consistent across different conditions, where a smaller proportion of lineage-specific genes were expressed compared to the rest of genes in K96243 genome (**Fig5 b to q**), suggesting that the lineage-specific genes were selectively expressed under specific conditions. Interestingly, there were a few conditions where the accessory genes specific to lineage 3 exhibited a higher number of gene expressions compared to compared to other genes. When *B. pseudomallei* was inoculated in distilled water for one hour, representing an early phase of nutrient deprivation, a greater proportion of accessory genes specific to lineage 3 (18 out of 29 genes, or 62.1%) were upregulated compared to the remaining genes in K96243 (2624 out of 5840, or 44.9%). Although this result only approached the borderline of significance (**Fig 5l**, p-val = 0.09), it is intriguing to observe an upregulation of lineage 3-specific genes under nutrient-depleted conditions. This finding suggests that lineage 3 may possess an adaptive strategy to persist in nutrient-depleted

soil, which is not uncommon in melioidosis endemic areas[25,26], before being acquired by a human host and subsequently causing the disease.

Example of lineage-specific genes

The majority of accessory genes specific to dominant lineages were located within the genomic islands (GI) (**Fig 5r** for lineage 3, p-val < 2.2×10^{-16}). These regions are characterised by anomalies in %G+C content or dinucleotide frequency signatures, or the presence of genes associated with mobile genetic elements such as insertion sequence (IS) elements and bacteriophages. Notably, we observed that a cluster of accessory genes specific to lineages 2 and 3 (*BPSS2060* to *BPSS2072*) formed a mosaic structure within a putative metabolic island known as GI 16[33]. Although several variations of GI 16 have been reported (GI16, GI16.1, GI16.2, GI16a, GI16b, and GI16b.1)[34], it typically spans 60 kb (*BPSS2051* to *BPSS2090*) and carries several known virulence determinants and genes that enhance metabolic versatility. While certain virulence factors, such as the filamentous hemagglutinin (*BPSS2053*) required for host cell adhesion and its processing protein were conserved across multiple lineages observed in our study, genes encoding functions that potentially expand the metabolic repertoire were specific to dominant lineages. For example, the mosaic structure of GI 16 (*BPSS2060* to *BPSS2072*), specific to lineages 2 and 3, contains genes involved in alternative nutrient catabolism and anabolism (*BPSS2060*, *BPSS2065*, *BPSS2067*, *BPSS2068*, and *BPSS2072*), a transcription regulator (B*PSS2061*), and substrate transport (*BPSS2064*, *BPSS2071*). Out of then 10 lineage-specific genes located in the mosaic structure of GI 16, seven were found to be upregulated during the early phase of nutrient starvation while remaining silent during infections. This finding reflects the functional division of GI16 where its lineage-specific mosaic structure contains genes that contribute to metabolic versatility, while its core structure encodes virulent determinants associated with disease implications. It is important to note that the structure of GI 16 may vary across different regions due to the plasticity of genomic islands and changes in selection pressures. While this observation is significant for northeast Thailand, it may not be generalisable to other geographical locations.

Lineage-specific genes were introduced by homologous recombination

Homologous recombination has been shown to play a significant role in facilitating the acquisition, loss, and introduction of mosaic structures within the GI of *B. pseudomallei*[35]. To better understand its association with lineage-specific genes, we quantified the rates of recombination in the dominant lineages. On average, the ratio of polymorphisms introduced

through recombination compared to those introduced by mutation (*r/m*) was 4.4 per lineage (**Table 1)**, with no observed differences in recombination rate between lineages (Kruskal-Wallis test, p-val = 0.11). When considering the combined genes from the four dominant lineages, 5,832 (89.5%) out of 6,519 genes had undergone recombination at least once (**Table 1; Supplementary fig. 4 and Supplementary data 6**). Within each lineage, recombination affected a large proportion of its genes. Specifically, the proportion of dominant accessory genes and genes carrying SNPs of “lineage 1”, “lineage 3”, “lineages 1 and 3” and “lineages 2 and 3” that were underwent recombination was 27/32 (79.4%), 339/383 (88.5%), 15/16 (93.8%) and 11/16 (68.8%), respectively (**Supplementary data. 7**). Notably, the restriction modification (RM) system, which is known to prevent the invasion of foreign DNA and restrict gene flow between *B. pseudomallei* lineages[36], was among the lineage-specific genes. A type III restriction system endonuclease (*BPSL0043*) which carries lineage-specific SNPs may act as a barrier for homologous recombination and potentially modulate lineage-specific genetic diversity. This highlights the intricate interplay between recombination, lineage-specific genes, and the RM system in shaping the genetic landscape of *B. pseudomallei* in northeast Thailand.

Discussion

Our analysis of bacterial population genomics enhanced our understanding of the evolution and adaptive strategies employed by the dominant clinical lineages of *B. pseudomallei* in northeast Thailand, a melioidosis region hyperendemic in Asia. Through an unprecedentedly dense sampling effort between 2015 and 2018, we were able to robustly characterise the coexistence of these dominant lineages, their unique genetic profiles, and the adaptive strategy of one lineage using comprehensive transcriptome data from a presentative strain K96243[31,32]. While our focus was on human clinical isolates, we anticipated that our findings are indicative of the broader environmental isolates as well. Previous studies have demonstrated a genetic overlap between clinical and environmental *B. pseudomallei* isolates in Ubon Ratchathani, a province located in northeast Thailand[27]. Additionally, smaller-scale investigations utilising MLST have also identified the prevalence of certain lineages in the environment of this region. This was exemplified by the presence of the previously identified ST 9, ST 48, ST 70, and ST 167, which were observed in environmental studies[27,37,38], and found to belong to clinical lineages 3, 1, and 4, respectively in our study. These findings suggest that a substantial portion, if not all, of the clinical and environmental isolates belong to the same lineages. Bacterial persistence in the environment serves as a prerequisite for host acquisition and subsequent infection. Therefore,

our observations on lineage-specific gene repertoire likely reflect selective advantages at various stages of the transition from the environment to host infection.

We hypothesised that the larger and persistent lineages over the sampling period could be considered successful, and their lineage-specific genetic content might provide insights into their adaptive strategies. Considering the wide range of niches occupied by *B. pseudomallei*, these strategies could facilitate bacterial survival during nutrient starvation in soil and water[21,24-26,39], as well as protect against predation by amoeba[40], or infection in animals and humans[13,41]. To mimic these diverse environmental and infection conditions, we performed a bacterial transcriptome analysis to identify the specific conditions or stages of the transition from the environment to host where lineage-specific genes were active. However, it is important to note that our analysis was based on a single reference strain from lineage 3. Given the observed population diversity, it is possible that other strains within lineage 3, as well as different lineages altogether, exhibit different expression patterns. While our interpretation is limited to the adaptive strategy of lineage 3, it still provides valuable insights into the conditions and lineage-specific genes that contribute to the success of this lineage.

Within lineage 3, we observed that the expression level of lineage-specific genes, including both accessory genes and genes carrying lineage specific SNPs, were generally lower compared to the remaining genes across the ranges of conditions we studied. This suggests that the lineage-specific genes were not constitutively expressed, but rather selectively upregulated under specific conditions. Notably, during an early phase of nutrient starvation, we observed a gene cluster on a putative metabolic island GI 16, specific to lineage 3, exhibited higher expression compared to the rest of the genome, albeit with borderline significance. The northeast region of Thailand is known for its poor soil quality, posing significant challenges for agriculture productivity and ecosystem health. This unfavorable environmental condition may have exerted selective pressure on *B. pseudomallei*, requiring the bacterium to adapt and survive as a prerequisite to infection. However, it is important to note that further validation is needed to establish a direct association between the lineage 3 - specific genes and bacterial survival under nutrient starvation, both *in vitro* and in the field. While our study revealed that most lineage-specific genes had undergone recombination and had the potential to be transferred between lineages, their gene flow might be restricted by lineage-specific restriction modification systems. Without conducting condition-specific expression experiments for each dominant lineage, it remains unclear whether the “persistence to infect later” strategy employed

by lineage 3 is adopted by other dominant lineages or if they employ different adaptive strategies.

Considering that host infection is believed to be an evolutionary dead end for *B. pseudomallei*, it is plausible that bacterial genetic factors contributing to lineage success may primarily enhance bacterial survival in environmental niches than during human infection. We demonstrated a high incidence of melioidosis during the rainy season and provided genetic evidence that the dispersal of *B. pseudomallei* is influenced by local climatic factors such as floods, seasonal monsoons and typhoons. Given the growing concerns regarding climate change, the reported poor soil nutrient conditions in the area, and the adaptive strategies of *B. pseudomallei* to persist in such environmental conditions; we hypothesis that the environmental conditions in northeast Thailand may provide an ideal incubator for the thriving of the currently dominant lineages of *B. pseudomallei.* Based on the significant role played by environmental conditions, disease control strategies should consider promoting both environmental and clinical public health.

Materials and Methods

Bacterial isolates

B. pseudomallei isolates used in this study comprised 1,265 clinical isolates collected from 1,265 melioidosis patients participated in our longitudinal cohort study conducted in northeast Thailand from July 2015 to December 2018[7]. The patients were from nine provinces, including Udon Thani (n = 230), Mukdahan (n = 198), Roi Et (n = 195), Surin (n = 170), Nakhon Phanom (n = 135), Buriram (n = 123), Sisaket (n = 107), Khon Kaen (n = 96) and Maha Sarakham (n = 11), who were admitted to nine hospitals included in our cohort (**Fig. 1a**). The ethical approval for the cohort study was obtained from the Ethics Committee of the Faculty of Tropical Medicine, Mahidol University, with approval number, MUTM 2015-002-001 and MUTM 2021-055-01. The isolates were obtained from various clinical samples, including blood (71.8%), pus (12.5%), sputum (9.9%), body fluid (3.5%), urine (1.9%) and tissue (0.4%) (**Supplementary data 1**). As latent infection accounts for < 5% of the cases, the majority of clinical cases likely directly acquired from the environment. The numbers of enrolled cases was lower at the beginning of the study due to the delayed sample collection in some study sites, resulting in inconsistent number of bacterial isolates used across different sites. When applicable, a permutation test was performed to ensure that an unequal number of isolates did not impact the temporal or spatial analysis.

DNA extraction and whole genome sequencing

Genomic DNA was extracted from *B. pseudomallei* using QIAamp DNA Mini Kit (Qiagen, Germany). All genomic DNA were processed for the 150-base-read library preparation and sequencing using Illumina HiSeq2000 system with 100-cycle paired-end runs at Wellcome Sanger Institute, Cambridge UK. An average of 80X read depth was achieved. Accession numbers and origins of all genomes were listed in **Supplementary data 1**. To control the potential contamination in each sample with other closely related species, we assigned taxonomic identity using Kraken v.1.1.1[42].

Genome assembly and mapping alignment

The *de novo* assembly of short read data was performed using Velvet v.1.2.10[43] followed by optimisation as previously described[13]. Short reads were also mapped against *B. pseudomallei* strain K96243 (accession numbers BX571965 and BX571966) and variants called using Snippy v.4.6.0 (https://github.com/tseemann/snippy). To avoid mapping errors and false SNPs, we filtered out SNPs covered by less than 10 reads and found in a frequency of less than 0.9.

Population structure analysis

Population structure of *B. pseudomallei* was analyzed by the combination of two independent approaches: (i) PopPUNK v.2.4.0[44] with the easy-run option, min-kmer 15 and max-kmer 31 on 1,265 assembled genomes and (ii) ML phylogeny of core-SNP alignment with IQ-TREE v.2.0.3[45]. IQ-TREE was run with a standard model selection with the best fit model being determined as TVM+F+ASC+R6. A 1,000 bootstrap support was performed, with 76.6% of internal nodes having ≥ 80% bootstrap support. Combination of PopPUNK cluster and monophyletic structure of phylogeny based on core-SNP alignment was used to assigned lineages of *B. pseudomallei* in northeast Thailand. This resulted in 101 lineages (clusters) which were classified into lineage 1 to 101 based on the number of genomes per lineages. Lineage 1 had the highest number of genomes. Lineage 62 to 101 only contain singletons (consisted only one genome). Our population structure analysis demonstrated four main lineages namely lineage 1 to lineage 4 (n = 693/1,265 genomes). We defined these four lineages as dominant lineage (**Fig. 1b and 2)**. For other lineages, we defined as intermediate lineage if they had number of genomes ≥ 10 and if < 10, we considered as minor lineages (**supplementary data 1; Fig. 1b and 2)**. Both intermediate and minor lineage is called non-dominant lineages for downstream analysis. Pairwise SNP distances were calculated using snp-dists (v.0.7.0) (https://github.com/tseemann/snp-dists) on core-SNP alignment to determine the genetic

distances between isolates. Phylogeny and pairwise SNP distance data were visualised using iTOL[46] ggplot2[47], respectively.

Multi-locus Sequence Typing (MLST)

Sequences of seven MLST loci were defined from 1,265 Thai assemble genomes using CLC Microbial genomic module of CLC genomic workbench (v.12.0) (CLC Bio-Qiagen, Aarhus, Denmark). Allele and ST were defined using MLST database (https://pubmlst.org/organisms/burkholderia-pseudomallei).

Specific lineage analysis

To investigate the dissemination and genetic diversity of each *B. pseudomallei* lineage, we performed a genome alignment, recombination removal and ML phylogeny construction of each dominant lineage individually. To increase variant calling sensitivity, a closely related genome was chosen for using as reference for each dominant lineage.

For lineage 1, a represent isolate DR10036A was selected and subjected for long-read sequencing on a MinION sequencer, according to the standard protocol provided by the manufacturer (Oxford Nanopore Technologies, Oxford, United Kingdom). A complete hybrid assembly of long-read and short-read sequence data of this strain was performed using Unicycler v.0.8.4[48]. For lineage 2 and 4 where representative complete genomes were not available, we selected a *de novo* assembly of one isolate within each of these lineages to assemble and order according to the complete genome of strain K96243 using ABACAS v.1.3.1[49], and then used them as reference for mapping. The complete genome *B. pseudomallei* strain K96243 (accession numbers BX571965 and BX571966) was used a reference sequence for lineage 3 because as it was assigned into this lineage by both PopPUNK and phylogenetic analysis.

We employed Snippy v.4.6.0 (https://github.com/tseemann/snippy) with option to filter out variants that were covered by less than 10 reads and found in a frequency of less than 0.9, to achieved genome alignment of each lineage. *Burkholderia thailandensis* genome strain Bp7055 (accession number SRR2972898) was used as the outgroup. All genome alignments were subjected to Gubbins v.3.1.3[50], to identify and remove recombination fragments. Identification of recombination of genome alignment was employed to address the genetic diversity within each dominant lineage and to produce recombination-free variant alignment for lineage-specific phylogeny reconstruction.

Lineage-specific phylogenetic analysis was then performed to investigate *B. pseudomallei* dissemination. Briefly, maximum-likelihood phylogenies were constructed on recombination-free variant alignment of each dominant lineage using IQ-TREE v.2.0.3[45] with general time reversible model with four gamma categories for rate heterogeneity and 1,000 replicates of bootstrap support. The overall proportion of nodes showing ≥80% bootstrap support of lineage-specific phylogenies was 68.7%.

Ancestral state reconstruction analysis

Ancestral state reconstruction was performed within each dominant lineage to understand the dissemination of *B. pseudomallei* among provinces in Northeast Thailand. Since the number of genomes from all 9 provinces are not consistent, province with small number (< 100 genomes) such as Khon Kaen (n = 96) and Maha Sarakham (n = 11) were excluded. Seven provinces including Buriram, Surin, Sisaket, Roi Et, Mukdahan, Nakhon Phanom and Udon Thani were selected for ancestral state reconstruction analysis. This resulted in 21 combinations of province-to-province transmission pairs. To reduce sampling bias, we sub-sampled the phylogeny of each dominant lineage to have an equal number of isolates per province (n = 15 isolates for lineage 1, n = 10 for lineage 2 and 4, and n = 7 for lineage 3) and permuted 1,000 times. We used stochastic character map (*make.simmap*) function of R package phytools v.4.0.5[51] with 100 simulations (nsim = 100) to determine the time of transitions across different provinces. Bidirectional transmission was defined as observed similar or different number of transition events of opposite direction, but inconsistent among different lineages. Unidirectional was defined as observed different number of transition events of opposite direction and the different number of events was consistent for all tested lineages. For unidirectional transmission, Mann-Whitney U test was used to compare the number of transition events between provinces.

Identification of dominant lineage-specific genes and SNPs

To identify genetic determinants contributed to the success of dominant population, we detected genes and SNPs that were presented in genomes of dominant lineages (n = 693 isolates), but likely absent in non-dominant lineages (n = 572 isolates). These genetic markers are called dominant lineage-specific genes and SNPs. To detect dominant lineage-specific genes, we first employed Prokka v.1.13.4[52] on 1,265 assemble genomes to annotate all genes using annotated *B. pseudomallei* K96243 genome as reference. We estimated the pan-genome employed using Panaroo v.1.2.9 with moderate mode, default core-genome sample threshold (0.95) and 95% sequence identity, the optimal threshold for our dataset (**Supplementary fig. 5**)

and a previous study[53]. This resulted in 12,734 pan-genes consisted of 5,514 core (cut-off ≥ 95%), and 7,220 accessory genes (cut-off < 95%) across the 1,265 samples used in this study, which is similar to data reported elsewhere. Since core genes and genes presented in low frequency may not specific to dominant lineages, only accessary genes presented in 5-94% of the total population were retained to detect dominant lineage-specific genes. This filter criteria resulted in 3,451 accessary genes for further analysis. Here, dominant lineage-specific genes were defined as genes present at ≥95% of isolates grouped in any of four dominant lineages, but present at ≤5% (strict threshold), ≤10% (intermediate threshold), or ≤15% (relaxed threshold) in isolates outside dominant lineages. We then detected dominant lineage-specific SNPs by reconstructing core-SNPs of all 1,265 genomes on the ML phylogeny using reconstruct_snps_on_tree.py script of BactGenScripts software (https://github.com/sanger-pathogens/bact-gen-scripts). Core-SNPs that belonged to internal nodes of each dominant lineage, but likely absent in non-dominant lineages were then defined as dominant lineage-specific SNPs. R version 4.0.4 was used to parse the data of this analysis.

Gene functional category

Biological function of dominant lineage-specific genes and genes carried dominant lineage-specific SNPs were annotated using Clusters of orthologous genes (COG) terms of *B. pseudomallei* K96243 available on the *Burkholderia* Genome Database[54]. Since the COG terms were not available for all *B. pseudomallei* CDSs, 69% (310/449 genes) of lineage-specific genes was assigned (**Supplementary data 2**). The association between COG terms and protein domain with lineage-specific genes and SNPs was performed using one-sided Fisher's exact test and corrected for multiple testing using Benjamini-Hochberg methods.

Transcriptomic analysis of dominant lineage-specific genes and SNPs

Expression condition was characterised based on the microarray expression data generated by Ooi et al.[31,32]. The analysis focused on the experiments performed on a reference strain K96243, a representative strain from lineage 3. A total of 65 conditions across both environmental and infection conditions were studied using the binary expression data. For each expression condition, the proportion of genes upregulation between each gene category including accessory genes specific to lineage 3, genes carrying SNPs specific to lineage 3, and the remaining genes in K96243 was compared. A one-sided Fisher's exact test was performed

on the number expressed genes of specific to lineage 3 vs the rest of the genes with higher expression in lineage-specific genes served as an alternative hypothesis. None of the studied conditions achieved significance following Benjamini-Hochberg correction for multiple tests.

Acknowledgements

We would like to thank physician and microbiology staff at the Udon Thani Hospital, Khon Kaen Hospital, Srinakarin Hospital, Nakhon Phanom Hospital, Mukdahan Hospital, Roi Et Hospital, Surin Hospital, Sisaket Hospital and Buriram Hospital for reporting the case and providing the facilities for bacterial collection. We are grateful to Dr. Soe Htet Aung for support for geographical map plot. This work was funded by Mahidol University (MU-KMUTT Biomedical engineering and Biomaterials Consortium) and Royal Golden Jubilee Ph.D. Programme (RGJ-ASEAN) (http://rgj.trf.or.th) through RS and NC. CC was funded by Wellcome International Intermediate Fellowship (216457/Z/19/Z) and Sanger International Fellowship. NC and TEW were supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (NIAID/NIH) (https://www.nih.gov) under Award Number U01AI115520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was funded in part, by the Wellcome Trust [220211]. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The genome sequence data presented in this study can be found in online repositories. The ENA under study accession number PRJEB25606 and PRJEB35787. The accession numbers for individual genomes listed in Supplementary data 1.

Author contributions

N.C., and C.C. and R.S. conceived, designed the study and write the original draft. N.C., R.S., T.E.W., and N.S. collected and identified bacterial isolates. N.C., R.S. and R.P. collected clinical data. C.C., N.C., R.S., N.R.T., J.T. and W.C. performed whole-genome sequencing. N.C. and C.C. contributed reagents. R.S., C.C., N.C., C.C., S.T. and E.B. performed bioinformatics

analysis. N.C. and C.C. administrated and supervised the project. All authors read and approved the manuscript.

Figure legend

Figure 1 Regional distribution of *B. pseudomallei* genomes. **(a)** Map of sampled provinces (based on patient's residence) for 1,265 Thai *B. pseudomallei* genomes. Circle size corresponds to the number of genomes. **(b)** Number of *B. pseudomallei* genomes in each dominant and intermediate lineage. **(c)** Pairwise SNP distance of within and between each dominant and intermediate lineage. **(d)** Temporal distribution of samples by lineage. Shaded area corresponds to the rainy season during May to October. Minor lineages are not included in this plot.

Figure 2 Maximum likelihood phylogeny of 1,265 Thai *B. pseudomallei* isolates. Colored strips show bacterial isolation location (province), year, ST and PopPUNK lineage. The tree was rooted on an Australian *B. pseudomallei* strain MSHR5619.

Figure 3 Dissemination of *B. pseudomallei* in Northeast Thailand. **(a)** Geographical distribution of the 4 dominant lineages. Spatial coordinates for each genome are the patient home address collected during enrollment. **(b)** Proportion of node transitions (%) between studied provinces. Red and blue spot are the median value. **(c)** Map showing geographical connectivity in Northeast Thailand. Colored arrows correspond to and extensive river system (lightblue), Northeast monsoon (pink), typhoon (red) and Mekong river flow (blue).

Figure 4 Dominant lineage-specific genes and SNPs. **(a)** Phylogeny (left), lineages (colored strips) and heat-map of lineage-specific genes (red) and SNPs (blue) presented in dominant lineages. **(b)** Functional categories of genes (COGs) of dominant lineage specific genes. **(c)** Functional categories of genes (COGs) of genes carrying dominant lineage specific SNPs.

Figure 5 The expression of K96243 genes across different environmental and infection conditions. Lineage 3 specific markers were mapped to the genome of K96243, enabling a comparison of gene expression across different conditions. The expression level of accessory genes specific to lineage 3 (red), genes carrying SNPs specific to lineage 3 (yellow) and the remaining of the genome were marked by different colour. **(a)** A boxplot summarizing the number of conditions each gene was expressed (a total of 65 conditions). The box and middle line represent first quantile, median, and third quartile, respectively. **(b)** to **(q)** the bars summarized the proportion of genes being expressed in each condition. The number of gene expression was compared for each gene categories. **(r)** A proportion of gene located in know genomic island (GI). Wilcox test was used to compare continuous variable, while the Fisher's exact test was employed to compare binary data

Table 1 Recombination among dominant lineages.

Lineage	Impacted CDS	Average *r/m*		
		Internal nodes (95% CI)	Terminal nodes (95% CI)	Average (95% CI)
1	5,859	6.6 (3.0-10.3)	4.8 (4.1-5.7)	5.7 (3.9-7.6)
2	5,733	1.8 (1.3-2.6)	5.4 (4.4-6.4)	3.6 (3.1-4.2)
3	5,947	4.1 (3.1-5.1)	5.1 (3.9-6.3)	4.6 (3.8-5.4)
4	5,787	1.4 (1.8-1.8)	3.4 (2.5-4.3)	2.4 (1.9-2.9)
Average	5,832	4.1 (2.6-5.5)	4.8 (4.3-5.3)	4.4 (3.7-5.2)

References

1 Limmathurotsakul, D. *et al.* Predicted global distribution of *Burkholderia pseudomallei* and burden of melioidosis. *Nat Microbiol* **1**, 15008 (2016).

2 Mohapatra, P. R. & Mishra, B. Burden of melioidosis in India and South Asia: Challenges and ways forward. *The Lancet Regional Health - Southeast Asia* **2**, doi:10.1016/j.lansea.2022.03.004 (2022).

3 Zheng, X., Xia, Q., Xia, L. & Li, W. Endemic Melioidosis in Southern China: Past and Present. *Trop Med Infect Dis* **4**, doi:10.3390/tropicalmed4010039 (2019).

4 Tran, Q. T. L. *et al.* Child Melioidosis Deaths Caused by *Burkholderia pseudomallei*-Contaminated Borehole Water, Vietnam, 2019. *Emerg Infect Dis* **28**, 1689-1693, doi:10.3201/eid2808.220113 (2022).

5 Win, M. M. *et al.* Enhanced melioidosis surveillance in patients attending four tertiary hospitals in Yangon, Myanmar. *Epidemiol Infect* **149**, 1-23, doi:10.1017/s095026882100128x (2021).

6 Thabit, A. *et al.* Etiologies of tropical acute febrile illness in West Pahang, Malaysia: A prospective observational study. *Asian Pacific Journal of Tropical Medicine* **13**, 115-122, doi:10.4103/1995-7645.278096 (2020).

7 Chantratita, N. *et al.* Characteristics and one year outcomes of melioidosis patients in Northeastern Thailand: a prospective, multicenter cohort study. *The Lancet Regional Health - Southeast Asia* **9**, doi:10.1016/j.lansea.2022.100118 (2023).

8 Hodgetts, K. *et al.* Melioidosis in the remote Katherine region of northern Australia. *PLOS Neglected Tropical Diseases* **16**, e0010486, doi:10.1371/journal.pntd.0010486 (2022).

9 Gassiep, I., Ganeshalingam, V., Chatfield, M. D., Harris, P. N. A. & Norton, R. E. The epidemiology of melioidosis in Townsville, Australia. *Trans R Soc Trop Med Hyg* **116**, 328-335, doi:10.1093/trstmh/trab125 (2022).

10 Bulterys, P. L. *et al.* Climatic drivers of melioidosis in Laos and Cambodia: a 16-year case series analysis. *Lancet Planet Health* **2**, e334-e343, doi:10.1016/s2542-5196(18)30172-4 (2018).

11 Chai, L. Y. A. & Fisher, D. Earth, wind, rain, and melioidosis. *The Lancet Planetary Health* **2**, e329-e330, doi:10.1016/S2542-5196(18)30165-7 (2018).

12 Liu, X. *et al.* Association of melioidosis incidence with rainfall and humidity, Singapore, 2003-2012. *Emerg Infect Dis* **21**, 159-162, doi:10.3201/eid2101.140042 (2015).

13 Chewapreecha, C. *et al.* Global and regional dissemination and evolution of *Burkholderia pseudomallei. Nature microbiology* **2**, 16263-16263 (2017).

14 Zheng, H. *et al.* Genetic diversity and transmission patterns of *Burkholderia pseudomallei* on Hainan island, China, revealed by a population genomics analysis. *Microb Genom* **7**, doi:10.1099/mgen.0.000659 (2021).

15 Sarovich, D. S. *et al.* Phylogenomic Analysis Reveals an Asian Origin for African *Burkholderia pseudomallei* and Further Supports Melioidosis Endemicity in Africa. *mSphere* **1**, e00089-00015, doi:doi:10.1128/mSphere.00089-15 (2016).

16 Price, E. P., Currie, B. J. & Sarovich, D. S. Genomic Insights Into the Melioidosis Pathogen, *Burkholderia pseudomallei. Current Tropical Medicine Reports* **4**, 95-102, doi:10.1007/s40475-017-0111-9 (2017).

561 17 Currie, B. J. *et al.* A cluster of melioidosis cases from an endemic region is clonal and is linked to
562 the water supply using molecular typing of *Burkholderia pseudomallei* isolates. *Am J Trop Med*
563 *Hyg* **65**, 177-179, doi:10.4269/ajtmh.2001.65.177 (2001).

564 18 Draper, A. D. *et al.* Association of the melioidosis agent *Burkholderia pseudomallei* with water
565 parameters in rural water supplies in Northern Australia. *Appl Environ Microbiol* **76**, 5305-5307,
566 doi:10.1128/aem.00287-10 (2010).

567 19 Chen, Y.-L. *et al.* The Concentrations of Ambient *Burkholderia Pseudomallei* during Typhoon
568 Season in Endemic Area of Melioidosis in Taiwan. *PLOS Neglected Tropical Diseases* **8**, e2877,
569 doi:10.1371/journal.pntd.0002877 (2014).

570 20 Chen, P.-S. *et al.* Airborne Transmission of Melioidosis to Humans from Environmental Aerosols
571 Contaminated with *B. pseudomallei. PLOS Neglected Tropical Diseases* **9**, e0003834,
572 doi:10.1371/journal.pntd.0003834 (2015).

573 21 Limmathurotsakul, D. *et al.* Melioidosis caused by *Burkholderia pseudomallei* in drinking water,
574 Thailand, 2012. *Emerg Infect Dis* **20**, 265-268 (2014).

575 22 Yip, T. W. *et al.* Endemic melioidosis in residents of desert region after atypically intense rainfall
576 in central Australia, 2011. *Emerg Infect Dis* **21**, 1038-1040, doi:10.3201/eid2106.141908 (2015).

577 23 Baker, A. L., Ezzahir, J., Gardiner, C., Shipton, W. & Warner, J. M. Environmental Attributes
578 Influencing the Distribution of *Burkholderia pseudomallei* in Northern Australia. *PLOS ONE* **10**,
579 e0138953, doi:10.1371/journal.pone.0138953 (2015).

580 24 Pumpuang, A. *et al.* Survival of *Burkholderia pseudomallei* in distilled water for 16 years. *Trans R*
581 *Soc Trop Med Hyg* **105**, 598-600, doi:10.1016/j.trstmh.2011.06.004 (2011).

582 25 Chewapreecha, C. *et al.* Co-evolutionary Signals Identify *Burkholderia pseudomallei* Survival
583 Strategies in a Hostile Environment. *Mol Biol Evol* **39**, doi:10.1093/molbev/msab306 (2022).

584 26 Hantrakun, V. *et al.* Soil Nutrient Depletion Is Associated with the Presence of *Burkholderia*
585 *pseudomallei. Appl Environ Microbiol* **82**, 7086-7092, doi:10.1128/aem.02538-16 (2016).

586 27 Chewapreecha, C. *et al.* Genetic variation associated with infection and the environment in the
587 accidental pathogen *Burkholderia pseudomallei. Communications Biology* **2**, 428,
588 doi:10.1038/s42003-019-0678-x (2019).

589 28 Floch, P. L. & Molle, F.

590 29 Arpornrat, T., Ratjiranukool, S., Ratjiranukool, P. & Sasaki, H. Evaluation of Southwest Monsoon
591 Change over Thailand by High-resolution Regional Climate Model under High RCP Emission
592 Scenario. *Journal of Physics: Conference Series* **1144**, 012112, doi:10.1088/1742-
593 6596/1144/1/012112 (2018).

594 30 Vongvisessomjai, S. Effect of global warming in Thailand. *Songklanakarin Journal of Science and*
595 *Technology* **32**, 431-444 (2010).

596 31 Ooi, W. F. *et al.* The condition-dependent transcriptional landscape of *Burkholderia*
597 *pseudomallei. PLoS Genet* **9**, e1003795, doi:10.1371/journal.pgen.1003795 (2013).

598 32 Heacock-Kang, Y. *et al.* The *Burkholderia pseudomallei* intracellular 'TRANSlTome'. *Nat Commun*
599 **12**, 1907, doi:10.1038/s41467-021-22169-1 (2021).

600 33 Holden, M. T. *et al.* Genomic plasticity of the causative agent of melioidosis, *Burkholderia*
601 *pseudomallei. Proc Natl Acad Sci U S A* **101**, 14240-14245 (2004).

602 603 34 Tuanyok, A. *et al.* Genomic islands from five strains of *Burkholderia pseudomallei. BMC Genomics* **9**, 566, doi:10.1186/1471-2164-9-566 (2008).

604 605 606 35 Spring-Pearson, S. M. *et al.* Pangenome Analysis of *Burkholderia pseudomallei:* Genome Evolution Preserves Gene Order despite High Recombination Rates. *PLoS One* **10**, e0140274, doi:10.1371/journal.pone.0140274 (2015).

607 608 609 36 Nandi, T. *et al. Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res* **25**, 129-141, doi:10.1101/gr.177543.114 (2015).

610 611 37 U'Ren J, M. *et al.* Fine-scale genetic diversity among *Burkholderia pseudomallei* soil isolates in northeast Thailand. *Appl Environ Microbiol* **73**, 6678-6681, doi:10.1128/aem.00986-07 (2007).

612 613 614 38 Vesaratchavest, M. *et al.* Nonrandom distribution of *Burkholderia pseudomallei* clones in relation to geographical location and virulence. *J Clin Microbiol* **44**, 2553-2557, doi:10.1128/jcm.00629-06 (2006).

615 616 39 Howard, K. & Inglis, T. J. Disinfection of *Burkholderia pseudomallei* in potable water. *Water Res* **39**, 1085-1092, doi:10.1016/j.watres.2004.12.028 (2005).

617 618 619 40 Noinarin, P., Chareonsudjai, P., Wangsomnuk, P., Wongratanacheewin, S. & Chareonsudjai, S. Environmental Free-Living Amoebae Isolated from Soil in Khon Kaen, Thailand, Antagonize *Burkholderia pseudomallei. PLOS ONE* **11**, e0167355, doi:10.1371/journal.pone.0167355 (2016).

620 621 622 41 Currie, B. J. *Burkholderia pseudomallei* and *Burkholderia mallei*: melioidosis and glanders. *Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases. 7th edn. Philadelphia: Churchill Livingstone Elsevier*, 2869-2885 (2010).

623 624 42 Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46, doi:10.1186/gb-2014-15-3-r46 (2014).

625 626 43 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).

627 628 44 Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* **29**, 304-316, doi:10.1101/gr.241455.118 (2019).

629 630 631 45 Trifinopoulos, J., Nguyen, L. T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* **44**, W232-235, doi:10.1093/nar/gkw256 (2016).

632 633 46 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293-W296, doi:10.1093/nar/gkab301 (2021).

634 635 47 Wickham, H. ggplot2. *WIREs Computational Statistics* **3**, 180-185, doi:https://doi.org/10.1002/wics.147 (2011).

636 637 638 48 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13**, e1005595, doi:10.1371/journal.pcbi.1005595 (2017).

639 640 641 49 Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968-1969, doi:10.1093/bioinformatics/btp347 (2009).

642 50 Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole
643 genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15, doi:10.1093/nar/gku1196 (2015).

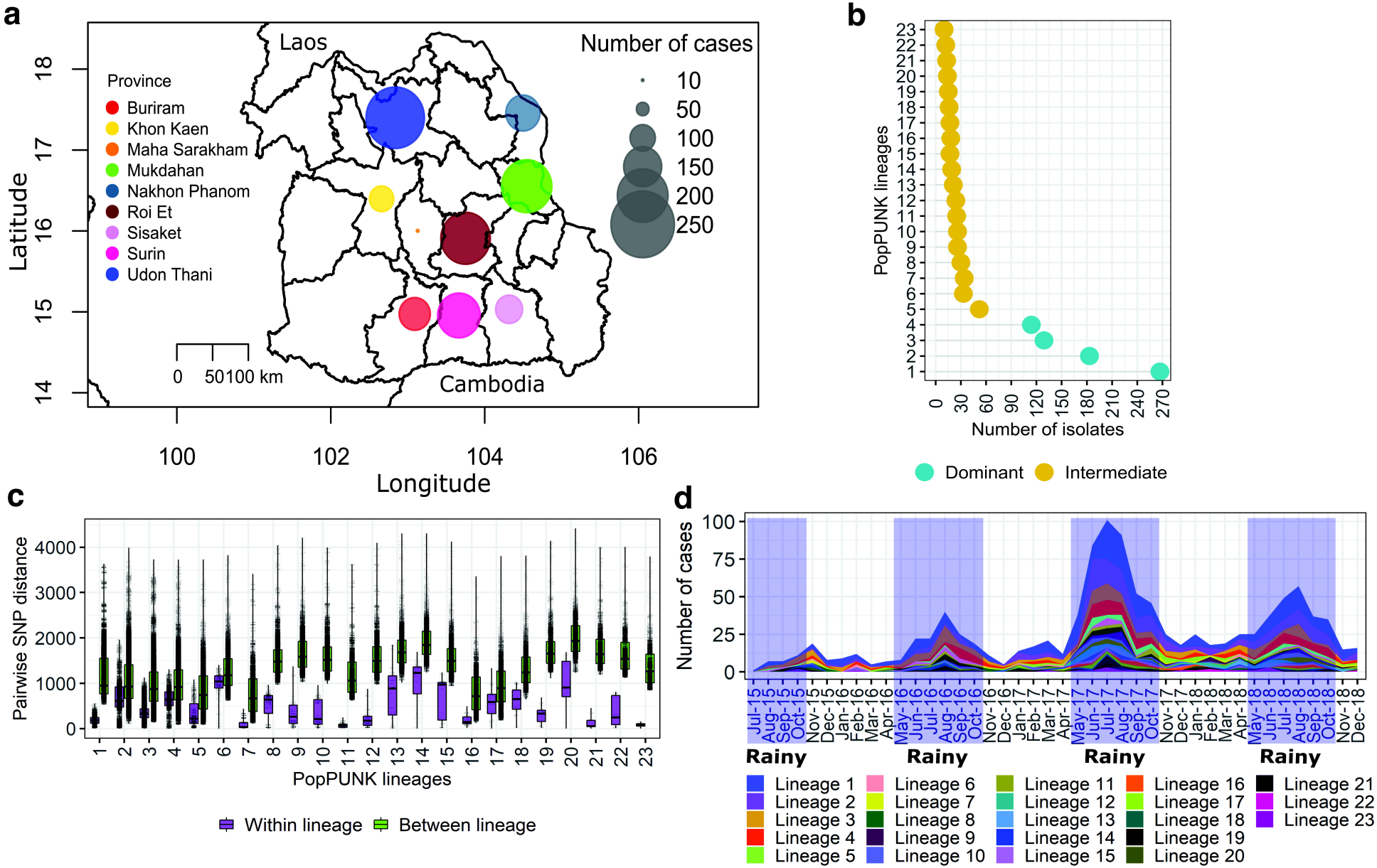
644 51 Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things).
645 *Methods in Ecology and Evolution* **3**, 217-223, doi:https://doi.org/10.1111/j.2041-
646 210X.2011.00169.x (2012).

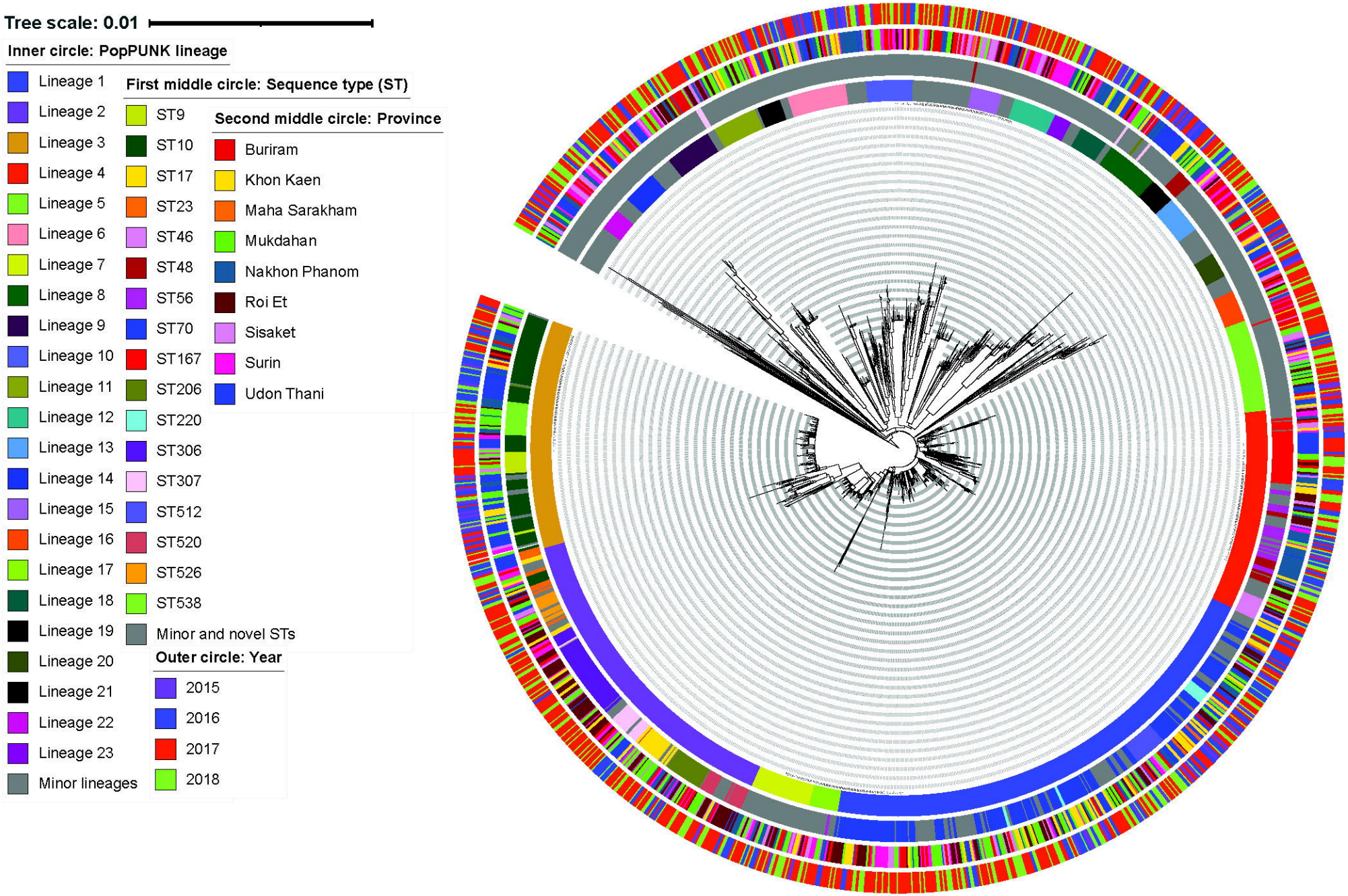
647 52 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069,
648 doi:10.1093/bioinformatics/btu153 (2014).

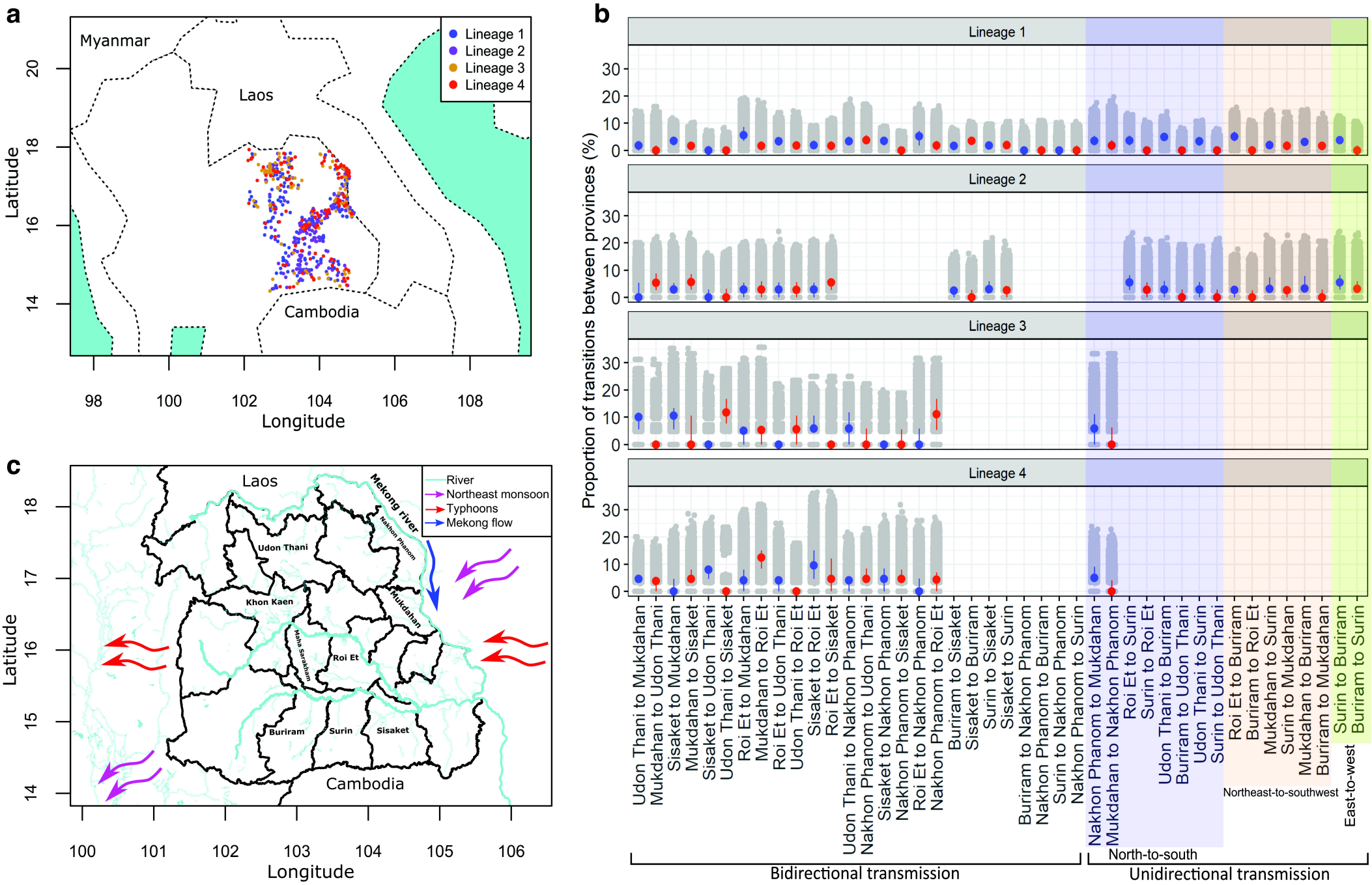
649 53 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI
650 analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**,
651 5114, doi:10.1038/s41467-018-07641-9 (2018).

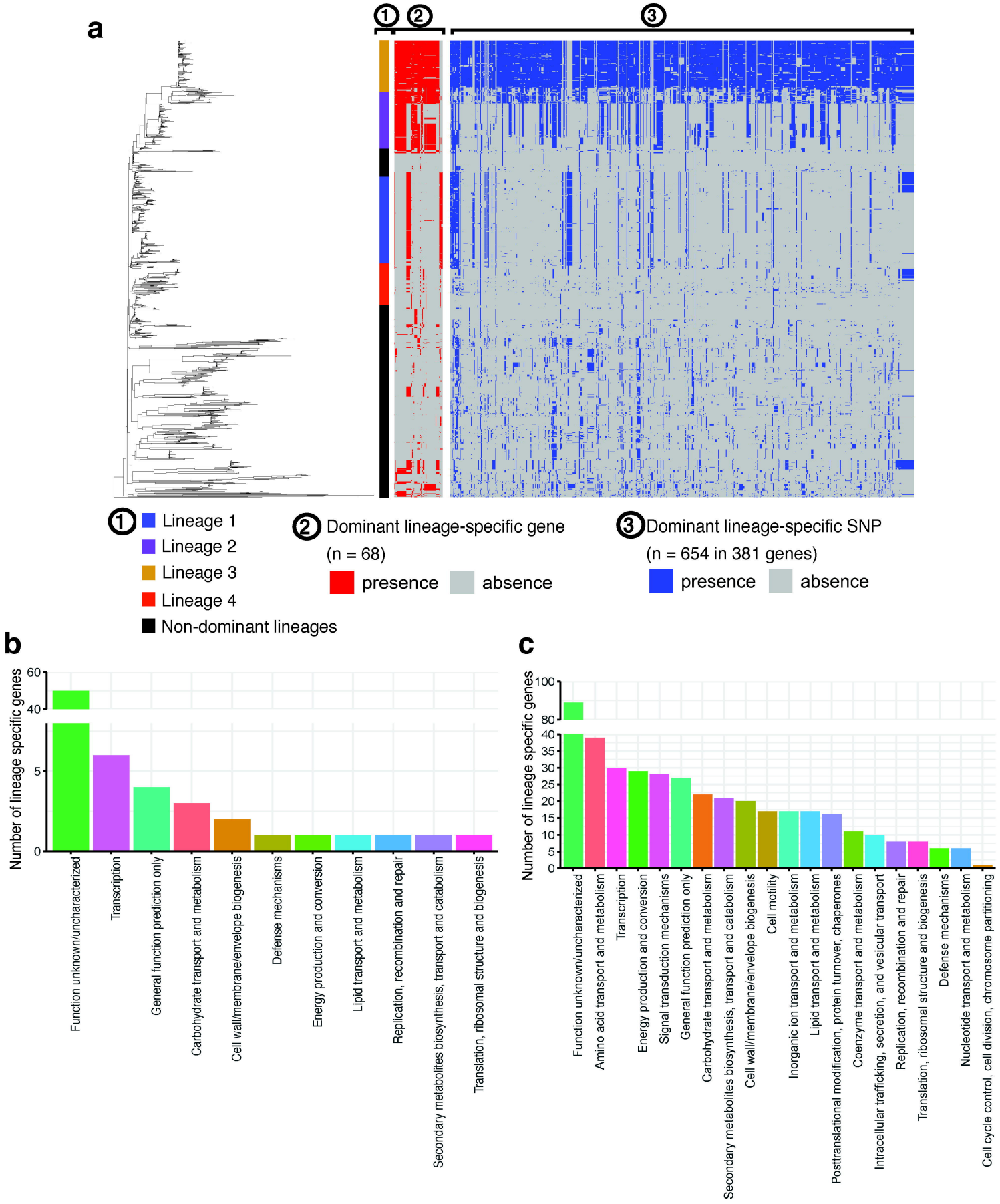
652 54 Winsor, G. L. *et al.* The *Burkholderia* Genome Database: facilitating flexible queries and
653 comparative analyses. *Bioinformatics* **24**, 2803-2804, doi:10.1093/bioinformatics/btn524 (2008).

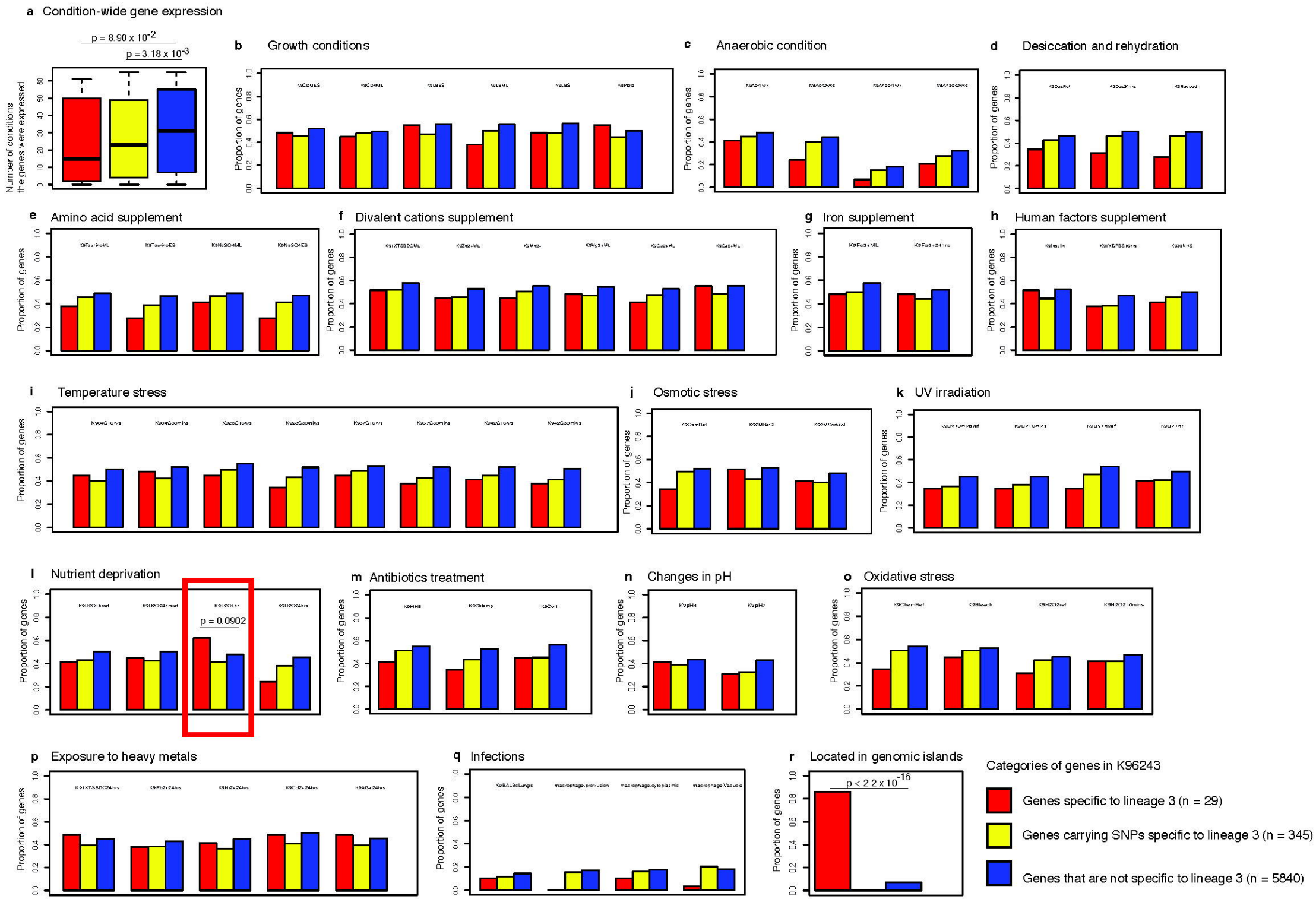
654


a
Laos
Cambodia
Number of cases
10
50
100
150
200
250
Province
Buriram
Khon Kaen
Maha Sarakham
Mukdahan
Nakhon Phanom
Roi Et
Sisaket
Surin
Udon Thani
0 50100 km
Latitude
Longitude
b
PopPUNK lineages
Number of isolates
Dominant
Intermediate
c
Pairwise SNP distance
PopPUNK lineages
Within lineage
Between lineage
d
Number of cases
Rainy
Rainy
Rainy
Rainy
Lineage 1
Lineage 2
Lineage 3
Lineage 4
Lineage 5
Lineage 6
Lineage 7
Lineage 8
Lineage 9
Lineage 10
Lineage 11
Lineage 12
Lineage 13
Lineage 14
Lineage 15
Lineage 16
Lineage 17
Lineage 18
Lineage 19
Lineage 20
Lineage 21
Lineage 22
Lineage 23


Tree scale: 0.01
Inner circle: PopPUNK lineage
Lineage 1
Lineage 2
Lineage 3
Lineage 4
Lineage 5
Lineage 6
Lineage 7
Lineage 8
Lineage 9
Lineage 10
Lineage 11
Lineage 12
Lineage 13
Lineage 14
Lineage 15
Lineage 16
Lineage 17
Lineage 18
Lineage 19
Lineage 20
Lineage 21
Lineage 22
Lineage 23
Minor lineages
First middle circle: Sequence type (ST)
ST9
ST10
ST17
ST23
ST46
ST48
ST56
ST70
ST167
ST206
ST220
ST306
ST307
ST512
ST520
ST526
ST538
Minor and novel STs
Second middle circle: Province
Buriram
Khon Kaen
Maha Sarakham
Mukdahan
Nakhon Phanom
Roi Et
Sisaket
Surin
Udon Thani
Outer circle: Year
2015
2016
2017
2018


a
Myanmar
Laos
Cambodia
Lineage 1
Lineage 2
Lineage 3
Lineage 4
Latitude
Longitude
b
Lineage 1
Lineage 2
Lineage 3
Lineage 4
Proportion of transitions between provinces (%)
Udon Thani to Mukdahan
Mukdahan to Udon Thani
Sisaket to Mukdahan
Mukdahan to Sisaket
Sisaket to Udon Thani
Udon Thani to Sisaket
Roi Et to Mukdahan
Mukdahan to Roi Et
Roi Et to Udon Thani
Udon Thani to Roi Et
Sisaket to Roi Et
Roi Et to Sisaket
Udon Thani to Nakhon Phanom
Nakhon Phanom to Udon Thani
Sisaket to Nakhon Phanom
Nakhon Phanom to Sisaket
Roi Et to Nakhon Phanom
Nakhon Phanom to Roi Et
Buriram to Sisaket
Sisaket to Buriram
Surin to Sisaket
Sisaket to Surin
Buriram to Nakhon Phanom
Nakhon Phanom to Buriram
Surin to Nakhon Phanom
Nakhon Phanom to Surin
Nakhon Phanom to Mukdahan
Mukdahan to Nakhon Phanom
Roi Et to Surin
Surin to Roi Et
Udon Thani to Buriram
Buriram to Udon Thani
Udon Thani to Surin
Surin to Udon Thani
Roi Et to Buriram
Buriram to Roi Et
Mukdahan to Surin
Surin to Mukdahan
Mukdahan to Buriram
Buriram to Mukdahan
Surin to Buriram
Buriram to Surin
Northeast-to-southwest
East-to-west
North-to-south
Bidirectional transmission
Unidirectional transmission
c
Laos
Cambodia
Mekong river
Nakhon Phanom
Udon Thani
Khon Kaen
Mukdahan
Maha Sarakham
Roi Et
Buriram
Surin
Sisaket
River
Northeast monsoon
Typhoons
Mekong flow
Latitude
Longitude


a
①
②
③
① Lineage 1
Lineage 2
Lineage 3
Lineage 4
Non-dominant lineages
② Dominant lineage-specific gene
(n = 68)
presence
absence
③ Dominant lineage-specific SNP
(n = 654 in 381 genes)
presence
absence
b
Number of lineage specific genes
60
40
5
0
Function unknown/uncharacterized
Transcription
General function prediction only
Carbohydrate transport and metabolism
Cell wall/membrane/envelope biogenesis
Defense mechanisms
Energy production and conversion
Lipid transport and metabolism
Replication, recombination and repair
Secondary metabolites biosynthesis, transport and catabolism
Translation, ribosomal structure and biogenesis
c
Number of lineage specific genes
100
80
40
35
30
25
20
15
10
5
0
Function unknown/uncharacterized
Amino acid transport and metabolism
Transcription
Energy production and conversion
Signal transduction mechanisms
General function prediction only
Carbohydrate transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
Cell wall/membrane/envelope biogenesis
Cell motility
Inorganic ion transport and metabolism
Lipid transport and metabolism
Posttranslational modification, protein turnover, chaperones
Coenzyme transport and metabolism
Intracellular trafficking, secretion, and vesicular transport
Replication, recombination and repair
Translation, ribosomal structure and biogenesis
Defense mechanisms
Nucleotide transport and metabolism
Cell cycle control, cell division, chromosome partitioning


a Condition-wide gene expression
p = 8.90 x 10^-2
p = 3.18 x 10^-3
Number of conditions the genes were expressed
b Growth conditions
c Anaerobic condition
d Desiccation and rehydration
e Amino acid supplement
f Divalent cations supplement
g Iron supplement
h Human factors supplement
i Temperature stress
j Osmotic stress
k UV irradiation
l Nutrient deprivation
p = 0.0902
m Antibiotics treatment
n Changes in pH
o Oxidative stress
p Exposure to heavy metals
q Infections
r Located in genomic islands
p < 2.2 x 10^-16
Proportion of genes
Categories of genes in K96243
Genes specific to lineage 3 (n = 29)
Genes carrying SNPs specific to lineage 3 (n = 345)
Genes that are not specific to lineage 3 (n = 5840)