Co-evolutionary signals from *Burkholderia pseudomallei* population genomics highlight its survival strategy in a hostile environment

Claire Chewapreecha[1,2,3*], Johan Pensar[4,5*], Supaksorn Chattagul[6,7*], Maiju Pesonen[8], Apiwat Sangphukieo[1,2,3], Phumrapee Boonklang[1], Chotima Potisap[6], Sirikamon Koosakulnirand[9], Edward J Feil[10], Susanna Dunachie[1,11], Narisara Chantratita[1,9], Direk Limmathurotsakul[1,11], Sharon J Peacock[12], Nick PJ Day[1,11], Julian Parkhill[13], Nicholas R Thomson[2], Rasana W Sermswan[6,7], Jukka Corander[2,13, 14]

Affiliations:

[1]Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand

[2]Wellcome Sanger Institute, Hinxton, Cambridge, UK

[3]Bioinformatics & Systems Biology Program, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

[4]Deparment of Mathematics, Faculty of Mathematics and Natural Sciences, University of Oslo, Norway

[5]Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[6]Melioidosis Research Center, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[7]Department of Biochemistry, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[8]Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

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

[10]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK

[11]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK

[12]Department of Medicine, University of Cambridge, UK

[13]Department of Veterinary Medicine, University of Cambridge, UK

[14]Department of Biostatistics, University of Oslo, Norway

*Contribute equally

Correspondence**:** Claire Chewapreecha (claire@tropmedres.ac), Rasana W Sermswan (rasana@kku.ac.th) & Jukka Corander (jukka.corander@medisin.uio.no)

The soil bacterium *Burkholderia pseudomallei* is the causative agent of melioidosis and a significant cause of human morbidity and mortality in many tropical and sub-tropical countries. The species notoriously survives harsh environmental conditions but the genetic architecture for these adaptations remains unclear. Here we employed a powerful combination of genome-wide epistasis and co-selection studies, gene expression analysis, and gene knockout assays to uncover genetic markers that have been repeatedly selected through *B. pseudomallei* evolution. These enabled us to identify a large number of mutation pairs under co-selection in distinct genes and non-coding RNA, of which a majority (61.5%) of signals detected in the discovery dataset can be replicated in the validation dataset (totalling 2,011 genomes). Genes under co-selection are mostly conditionally expressed with marked expression correlation of gene-gene pairs found in physical stress conditions. We discovered a putative adhesin (*BPSL1661*) as a hub of co-selection signals and experimentally confirmed the essential role of *BPSL1661* under nutrient deprivation. The gene co-selection network surrounding *BPSL1661* likely offers *B. pseudomallei* a selective advantage to survive nutrient limited conditions. We propose that anthropogenic activities such as pre- and post-harvest crop residue burning have accelerated soil nutrient depletion and may have directly contributed to the preferential survival of *B. pseudomallei* over benign soil microorganisms. This is expected to lead to a consequent increase in the incidence of melioidosis should the "slash-and-burn" practices continue to expand.

Main text

Introduction

The environmental bacterium *Burkholderia pseudomallei* has been increasingly recognised as an emerging human pathogen and a cause of melioidosis, a rapidly fatal infectious disease that annually affects and kills an estimated number of 165,000 and 89,000 patients, respectively[1]. Through direct environmental exposure, humans can acquire *B. pseudomallei* and develop melioidosis. The bacterium can be isolated from the soil in many tropical and sub-tropical regions. It is most abundant at depths ≥ 10 cm from the surface but can move from deeper soil layers to the soil surface during the rainy season and multiply; this has been linked with an increase in disease incidence after heavy rainfalls[2,3]. The presence of *B. pseudomallei* has been associated with nutrient-depleted soil and soil modified by long-term human activity[4–9]; thereby connecting human interference to soil physiochemistry that promotes the abundance of this species. *B. pseudomallei* can survive extreme environmental conditions ranging from dry terrain in deserts[10] to distilled water[11] (no nutrients); the latter experiment has been run for 25 years and is still ongoing (Dr Vanaporn Wuthiekanun, *personal communication*). Such extreme environmental conditions likely impose unique selection pressures on *B. pseudomallei*, and the identification of the genetic factors that mediate these adaptations has been a long-standing research objective. However, progress towards a better understanding of the evolution and adaptation of this species has been impeded by its large and complex gene repertoire. *B. pseudomallei* has at least two large chromosomes with a combined median size of 7.2 Mb, and a median of 5,936 genes with multiple genomic islands and paralogous genes[12–15]. A hypothesis-free co-selection analysis, known as genome-wide epistasis and co-selection study (GWES) has recently been used to successfully identify bacterial co-selected loci, many of which matched known gene-gene interacting partners[16,17]. This prompted us to apply GWES to detect co-evolutionary signals in a collection of 1,136 *B. pseudomallei* isolates and validate our findings on an additional collection of 875 isolates (**Extended Data Fig 1**, **Supplementary table 1**), enabling us to identify loci under recent selective pressure while distinguishing interacting genes from their multiple paralogs. To elucidate the function of the co-selected genes, we integrated the GWES findings with transcriptomic data[14,18] and gene knockout assays.

Results

Detection of co-evolutionary signals

Short reads from whole genome sequencing of the isolates in both the discovery and validation collection (**Supplementary Table 1**) were mapped against the *Burkholderia pseudomallei* K96243 reference genome with 389,476 single nucleotide polymorphisms (SNPs) being called (see methods). By performing GWES on the mapped sequence alignment, we discovered co-selection signals in 13,061 SNP-SNP pairs spanning chromosome I (5,550 pairs), chromosome II (7,309 pairs), and between chromosome I and II (202 pairs); of which 8,035 pairs (61.5%) could be replicated in the validation dataset. The congruence of the co-selection signals in the discovery and validation datasets is shown in **Supplementary table 2**, demonstrating that these complex patterns are common even across different datasets. The co-selected SNPs spanned genes, non-coding RNA (ncRNA) and other intergenic regions, each of which accounted for 69.5%, 6.8%, and 23.7% of the total signals, respectively. The majority of the pairs were either located within the same gene or in close physical proximity (**Extended Data Fig 2**); an observation driven by linkage disequilibrium (LD). These pairs were separated by small physical distances and are unlikely to have had recombination events between them, resulting in so-called LD-mediated links. The LD decayed when the distance between the pair increased, and most SNPs were approximately in linkage equilibrium when each site was on a separate chromosome. While pairs located in close proximity are most likely explained by LD structure, some of them may represent genuine *cis* interacting partners. These may include SNPs in regulatory regions upstream of genes, such as DNA-binding sites for RNA polymerase and transcription factors[19]; and ncRNA which predominantly acts in *cis*[18,20]. We thus grouped co-selected pairs into *cis*, and *trans* based on the length of transcription fragments reported in *B. pseudomallei*[14]. Here, pairs located on different genes or ncRNA but located within 7.68 kb (a size covering 95th percentile of polycistronic mRNAs) were termed *cis* pairs, while pairs located further than 7.68 kb apart or on separate chromosomes were termed *trans* pairs. This resulted in 334 *cis*- and 252 *trans* gene-gene pairs, 64 *cis*- and 11 *trans* gene-ncRNA pairs, and 19 *cis* ncRNA-ncRNA pairs from the discovery data; of which 383 pairs (56.3%) were replicated in the validation data (**Supplementary table 3**). Despite *cis* pairs being more prevalent, the number of interacting partners per gene is fewer than the number of *trans* interacting partners (a mean of partners per *cis*-gene = 1.11, a mean of partners per *trans*-gene = 2.50, with the highest number observed in *BPSL1661* (*trans*-gene partners = 104 (discovery data) and 52 (validation data), **Fig 1**). Although the *cis* interactions detected here are obscured by the LD structure, our observation is consistent with the idea of "*cis* dominates but *trans* is dominant": a gene expression concept suggesting that the majority of local regulatory variation functions in *cis*. However, *trans* loci commonly acts as "hotspots"

that affect large number of transcripts and predominantly act in a condition specific manner.[21–24]

Functional conservation of co-selected gene-gene pairs

We assigned functional annotations[25] to genes identified in both *cis* and *trans* gene-gene pairs. For both the discovery and validation dataset, the majority of the co-selected genes were functionally assigned as cell envelope (21.9%), or transport/binding protein (10.6%), while many were left as uncharacterised or hypothetical proteins (26.7%). Although there were moderate enrichments of terms associated with secondary metabolism, laterally acquired elements, and small molecule degradation in genes under co-selection (Benjamini-Hochberg adjusted p-value = 1.41×10^{-2}, 1.72×10^{-2}, and 3.81×10^{-2}, respectively, **Extended Data Fig 3a**), none of these reached the significance threshold across both the discovery and validation datasets. We observed that a high proportion of the co-selected gene-gene pairs share the same functional annotations (51.3% of *cis*- and 19.9 % *trans* gene-gene pairs). To test whether this observation was driven by chance, we generated a separate randomised dataset for "*cis*" and "*trans*" gene-gene pairs using the same distance threshold (n = 1,000 pairs each) and compared the number of pairs under the same or different functional annotations between the randomised and real datasets. Excluding pairs with ambiguous annotations, functional conservation of *cis* gene-gene pairs followed random expectation (**Extended Data Fig 3b**). A possible explanation is that bacterial genes are organised into operons to maximise efficiency in protein production and function[26], thereby increasing the chance of genes located in close physical proximity sharing the same functional annotation. However, functional conservation of *trans* gene-gene pairs was not at random (**Extended Data Fig 3c**, two-sided Fisher exact test: $trans_{\text{discovery vs randomised}}$ p-val = 1.35×10^{-4}, $trans_{\text{validation vs randomised}}$ p-val = 1.38×10^{-2}). Our results indicate that co-selection of genes, at least in *trans*, is driven by their functions and further supports the co-evolution theory[27] of the two chromosomes.

Expression patterns of co-selected gene-gene pairs

We explored the expression profile of co-selected gene-gene pairs using *B. pseudomallei* whole genome tiling microarray expression data generated by Ooi *et al*[14]. This covered a broad spectrum of conditions and exposure, including general growth (32 conditions), exposure to stress (33 conditions), infection (4 conditions), and genetic mutants (13 conditions) (**Extended Data Fig1c**). Here, genes expressed in ≥ 70 conditions[14] are defined as constitutively expressed as in Ooi *et al* [14]. Approximately 22.4% of the genes that were detected in co-selection pairs in the discovery analysis, validation analysis, or in both were constitutively expressed compared to 39.5% of genes that were not part of any co-

selection pairs (two-tailed Fisher's exact test $p = 5.74 \times 10^{-7}$), indicating that co-selection signals were more focused on condition-specific genes than on constitutively expressed genes.

We tested if *cis*- and *trans* gene-gene pairs demonstrating co-selection also showed expression correlation compared to any randomised *cis*- and *trans* gene-gene pairs (n random pairs = 1,000 each). In addition, we examined under which conditions such correlations were observed. Using normalised gene expression profiles[14], Pearson's correlation analysis was performed for each gene-gene pair with transcription data from all conditions as well as from subsets of conditions. For all conditions and 4/5 tested subsets of conditions, co-selected genes paired in *cis* showed greater expression correlation than randomly paired *cis*-genes (**Fig 2a,** two-sided Mann-Whitney U test p values $\leq 1.02 \times 10^{-2}$). The result suggests that expression of *cis* gene-gene pairs under co-selection is co-regulated. This is possibly mediated by the bacterial operon structure which enables the coupling of DNA transcription and messenger RNA translation, thereby enhancing the efficiency of protein production under each subset of conditions[28]. An absence of operon structure for distally located genes should lead to less expression correlation for *trans* gene-gene pairs. Nevertheless, we observed greater expression correlation of co-selected genes paired in *trans* than randomised controls under physical stress conditions (**Fig 2b**, two-sided Mann Whitney U test p values $\leq 1.05 \times 10^{-2}$) but less correlation under normal growth conditions (two-sided Mann Whitney U test p values $\leq 1.46 \times 10^{-2}$); suggesting that physical stress has largely shaped the *tran*s co-selection patterns at the time-scale considered for this population. Here, physical stress conditions included temperature stress, osmotic stress, UV irradiation and nutrient deprivation; the environmental conditions *B. pseudomallei* is regularly or seasonally exposed to[4–8,29]. However, it should be cautioned that *trans* gene-gene pairs are heavily driven by a few genes linked to many other genes. The expression profile or function of these few genes may thus largely drive the results observed here.

***BPSL1661* is a hotspot for co-selection**

Intriguingly, a putative adhesin gene *BPSL1661* constitutes 41.3% and 38.5% of *trans* gene-gene pairs in the discovery and validation datasets, respectively (**Fig 1b & c**). *BPSL1661* codes for an extracellular protein with a size ranging from 2,036 to 3,230 amino acids (approximately 325 kDa) in different isolates. The gene is located in a large highly variable genomic region termed genomic island 8 (GI8)[30], proposed to be acquired through horizontal gene transfer[27]. This genetic mobility possibly contributes to the presence of multiple *BPSL1661* alleles observed in our study. We detected 6 major alleles of *BPSL1661* ($n \geq 5$ isolates) consistent with previous reports on multiple protein epitopes[31,32] (**Extended Data Fig 4 & 5**). All *BPSL1661* alleles share a conserved hemolysin-type calcium binding

domain which is common in proteins secreted through a type I secretion system, and two copies of the VCBS domain (a repeat domain in *Vibrio*, *Colwellia*, *Bradyrhizobium* and *Shewanella*) which functions in cell adhesion. Allelic variations were observed in the presence of bacterial Ig and flagellin domains. Co-selection signals on *BPSL1661* did not scatter throughout the entire gene, but were focused on 8 SNPs, one of which was a non-synonymous V1701F mutation. None of these were located in predicted protein domains, and whether these have any impact on protein structure will be the subject of future investigation. Interestingly, previous studies reported heterogeneity in human immune responses to polypeptides generated from different *BPSL1661* alleles, ranging from null to strong antibody responses[31–33]. Such disparity in host recognition of different *BPSL1661* alleles potentially suggests that the protein may not principally function in host cell invasion but play other significant roles in *B. pseudomallei* survival. Transcription assays[14] further revealed that *BPSL1661* is down-regulated during infection but upregulated in acidic conditions (pH 4), mid-logarithmic phase in minimal media, and nutrient deprivation (**Fig 3a**); further indicating a role of *BPSL1661* in adaptation to environmental stress rather than host infection.

***BPSL1661* mediates bacterial survival under nutrient deprivation**

To better understand the function of *BPSL1661*, we knocked out the *BPSL1661* gene in the K96243 strain and compared the number of live cells of the mutant against the wildtype under the conditions *BPSL1661* was upregulated in[14] for 120 hours (**Fig 3b**). We observed no significant differences in bacterial survival in normal growth media, neutral pH (pH 7.4) or acidic conditions (pH 4). This could be due to functional redundancy that compensates for the loss of a single gene function. However, the *BPSL1661* mutant showed a significant reduction in stationary-phase survival compared with the wildtype under nutrient depleted condition (two-sided Kolmogorov-Smirnov test p-value = 3.73×10^{-5} and 3.25×10^{-3} for high and low bacterial inoculum, respectively); confirming an essential role of *BPSL1661* under nutrient deprivation. Our observation of *BPSL1661* as a hotspot may imply that nutrient depletion is the selective pressure underlying the co-selection patterns. Indeed, environmental sampling studies in Southeast Asia and Australia have consistently reported that the bacterium is commonly found in soils with lower nutrient availability[5,6]. Although lower nutrient abundance appears to be a common feature across melioidosis endemic areas; the soil physiochemical properties, microbial diversity, temporal disturbances such as monsoon seasons and anthropogenic activities that alter the environmental conditions vary greatly between each area and between studies [7,34–37]. These factors create patterns of spatial and temporal heterogeneity to which *B. pseudomallei* has adapted and possibly has led to the coexistence of multiple *BPSL1661* alleles detected in this study. We also noted

geographical differences in *BPSL1661* allele frequencies. An allele harbouring a flagellin domain (here denoted as allele A, **Extended Data Fig 4a**) was detected at lower frequency in Australia (28.7%), at moderate frequency in the Malay Peninsula (37.6% from Malaysia & Singapore) and higher frequency in the countries bordered by the great Mekong river (59.1 % from Thailand, Laos, Cambodia and Vietnam)(**Extended Data Fig 4b**). Such a difference in allele frequencies could be either driven by different local selection pressures, or by genetic drift (or both). Due to its horizontal mode of inheritance, an ancestral history of each *BPSL1661* allele could not be reliably reconstructed. Functional characterisation of different *BPSL1661* alleles also warrants further future studies.

***BPSL1661* co-selection network and putative bacterial response under low nutrient abundance**

We considered genes and ncRNA co-selected with *BPSL1661* in both the discovery (n gene pairs = 105, n ncRNA pairs = 5) and validation dataset (n gene pairs = 53, n ncRNA pairs = 2) totalling 136 pairs, of which 29 pairs are shared in both datasets (**Fig 4**, **Supplementary table 4**). The majority of these gene and ncRNA pairs were linked to *BPSL1661* in *trans* except for an outer membrane protein *BPSL1660* which paired in *cis*. During nutrient depleted conditions[14,18], 43/129 of the *trans* gene pairs and 1/6 of ncRNA pairs were upregulated with *BPSL1661*, while 64/129 of the *trans* gene pairs and 2/6 of ncRNA pairs were down-regulated (**Fig 4c**), respectively. Many of these genes are predicted to participate in alternative metabolic pathways, energy conservation, uptake of external carbon source, cellular signalling, and transcription regulation (**Supplementary table 4**).

A pyrophosphohydrolase (*SpoT* or *BPSL2561*) is upregulated during nutrient starvation and is known to have a dual function to synthesise and hydrolyse guanosine tetra- and pentaphosphates (ppGpp)[38]. In bacteria, ppGpp serves as a mediator in nutritional surveillance, coordinating a variety of cellular activities in response to changes in nutrient availability[39]. In *B. pseudomallei*, deletion of ppGpp synthetase and hydrolase led to reduced survival during stationary phase compared to wildtype[38]. It is possible that *B. pseudomallei* switches to alternative carbon sources to maintain cellular energy when complex sources such as glucose are not available. A C4-carboxylate transport transcription regulation protein (*BPSL0427*), a malate synthase (*BPSL2192*) and a glycogen branching enzyme (*BPSL2076*) were co-selected with *BPSL1661* but had different expression profiles under nutrient limited conditions. A transport transcription regulation protein (*BPSL0427*) is upregulated under nutrient deprivation. Its homolog was shown to facilitate the cellular uptake of four carbon compounds such as aspartate, fumarate, and succinate when common carbon sources such as glucose are scarce[40]. Another upregulated gene, malate synthase (*BPSL2192*) is a key enzyme involved in the glyoxylate cycle and was also shown

to be essential to *Mycobacterium tuberculosis* survival under nutrient starvation [41]. Its expression allows cells to utilise two-carbon compounds to sustain carbon requirement in the absence of glucose. On the contrary, a glycogen branching enzyme (*BPSL2076*) is co-selected with *BPSL1661* but downregulated during nutrient depletion. A homolog of glycogen branching enzyme is known to facilitate glucose conversion into long term glycogen storage when there are excessive carbon sources[42]. The selection of *SpoT*, *BPSL0427, BPSL2192* and *BPSL2076* through *BPSL1661* reflect the energy balance of cells under ranges of nutritional conditions through their evolutionary timeline. We also observed many ncRNA, transcription regulators and DNA binding proteins co-selected with *BPSL1661* (**Supplementary table 4**). These genes and ncRNA may further regulate downstream responses under changes in nutrient abundance. Together, the *BPSL1661* co-selection network seems to suggest that *B. pseudomallei* has adapted to survive nutrient limited conditions and hostile environments.

Discussion

This study is the first, in our knowledge, to deploy an integrated approach of GWES, transcriptomic analyses and knockout assays to understand the evolution and unique selective pressures acting on a micro-organism. While GWES detected signals for co-selected loci, transcriptomic data provided condition-dependent information on which selection pressures may have acted upon the detected loci. However, our study focused only on nucleotide polymorphisms found by comparison to the *B. pseudomallei* K96243 reference genome. Co-selected loci on other types of structural variants including indels, genomic inversions, gene duplication or horizontally acquired genes absent in K96342 will be missed from this analysis. It is also possible that other transcriptional conditions the bacterium is exposed to in its native niche are missing. The incomplete genomic and transcriptomic data warrant further studies to cover more complex genomic variants, and broader transcriptional conditions. Nevertheless, our integration of data offers stringent predictions on which genes are key to *B. pseudomallei* survival under specific conditions. In particular, the putative adhesin *BPSL1661* was identified as a hotspot of the co-selection map. Our gene knockout experiment confirmed that the gene is essential for survival under nutrient deprivation. This is consistent with the soil conditions in which *B. pseudomallei* are commonly found and provides evidence for the hypothesis that unlike most bacteria, *B. pseudomallei* has a competitive advantage in nutrient depleted environments.

Fires, if not moving slowly or consuming large areas, typically lead to a loss of organic matter, nutrients and microorganisms from the top few millimetres or centimetres of soil from the surface[43–45]. Given that *B. pseudomallei* typically colonises a deeper soil level,

fires are likely harmless to this species, but instead improve conditions for its survival. In melioidosis endemic areas, human-induced fires have been common for over a thousand years. In South- and Southeast Asia, crop residue burning after harvest is an ongoing practice and reported to link with reduced soil nutrients[46,47]. In Australia, bushfires and an ancient firestick farming practiced by Indigenous people over the past 100,000-120,000 years have largely shaped the continent's landscape and led to the spread of fire adapted species[48]. It is therefore notable that we observe genes required for survival of nutrient depletion at the hub of the co-selection signals. This predicts improved conditions for *B. pseudomallei* survival and a corresponding increase in the incidence of melioidosis should the "slash-and-burn" agricultural practices continue in these regions.

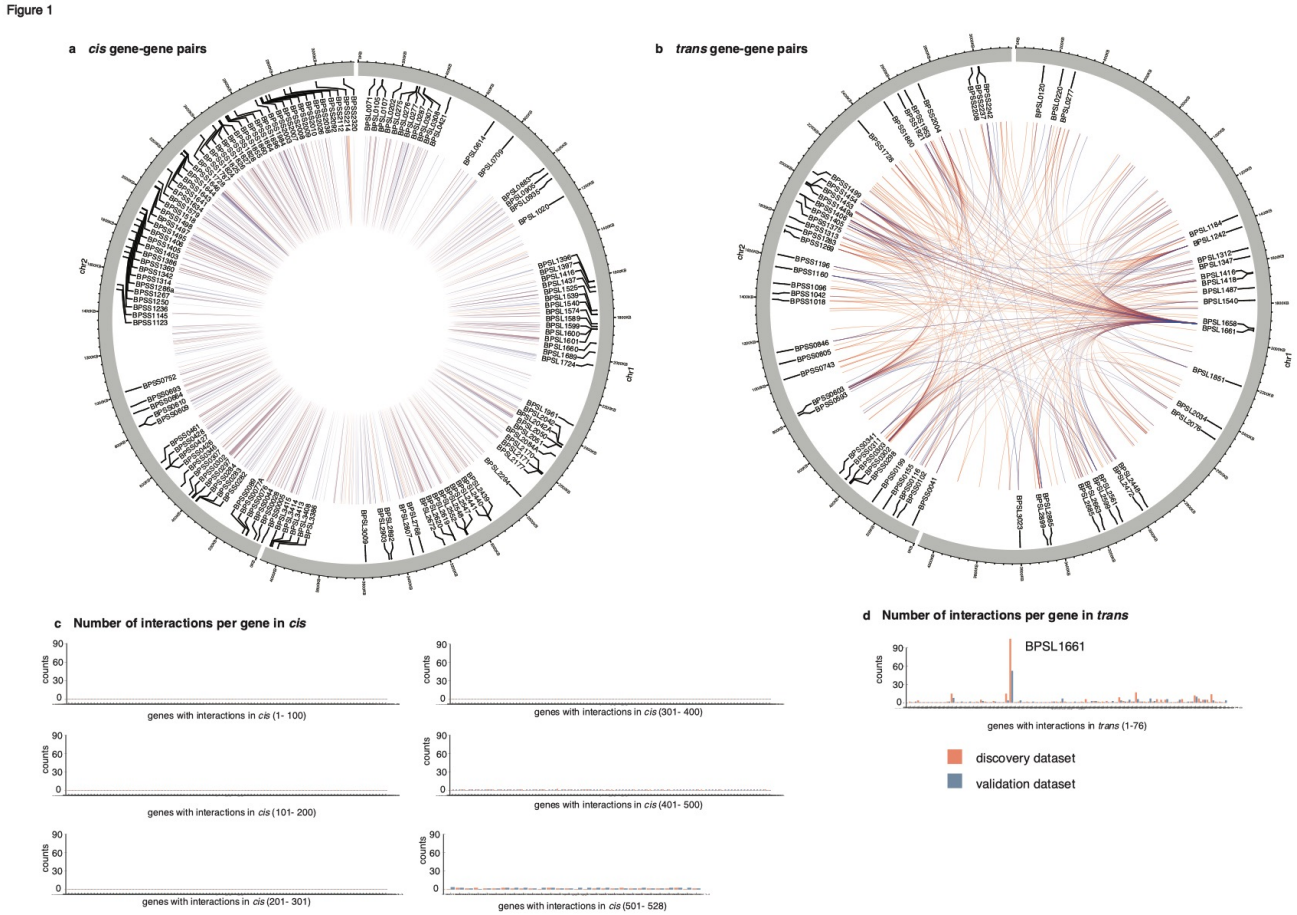


Fig 1 Co-selected gene-gene pairs in discovery and validation datasets. a, b) circos diagrams showing co-selected gene-gene pairs in *cis* (a) and *trans* (b). Any pairs located within 7.68 kb (95th percentile of transcription fragments) were categorised as *cis*, while those located further apart were grouped as *trans*. c, d) bar charts summarising the number of interacting partners per gene for *cis* (c) and *trans* (d) linkage. Data from discovery and validation datasets were highlighted in red and blue, respectively.

Figure 2

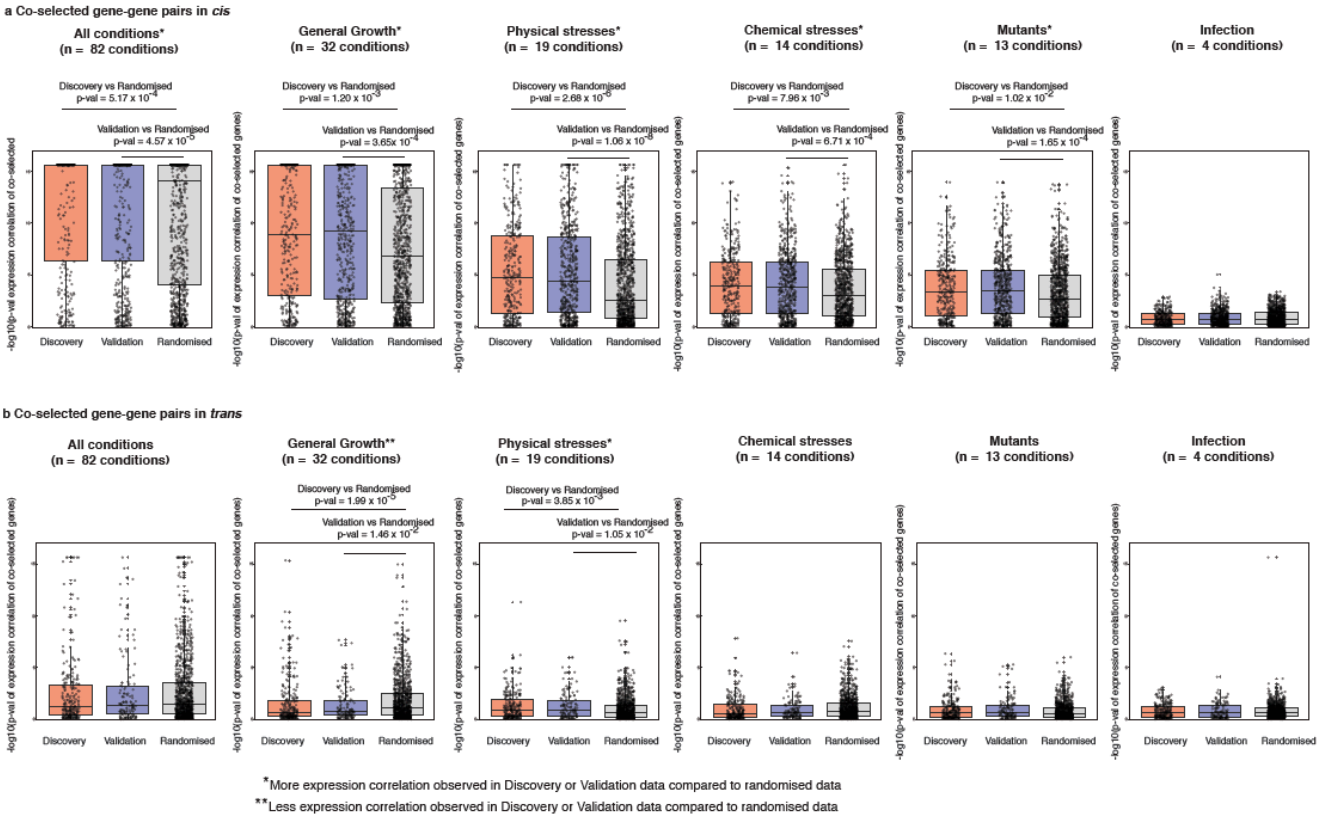


Fig 2 Expression correlation of co-selected gene-gene pairs. a) & b) represent expression correlations of co-selected gene-gene pairs from the *cis*- and *trans* interactions, respectively. For all plots, vertical axis denotes the strength of expression correlation on a negative log p-value scale while horizonal axis represents groups of data including true co-selection signals found in discovery (red), validation (blue) and randomised dataset. The distribution of significant expression correlation of either discovery or validation data was compared to randomised data using two-sided Mann Whitney U test. The expression conditions tested (from left to right) include all conditions (n =82), general growth (n=32), physical stresses (n=19), chemical stresses (n=14), mutants (n=13) and infection (n=4). Each boxplot summarises first quartile, median, and third quartile with individual value plotted as a dot.

Figure 3

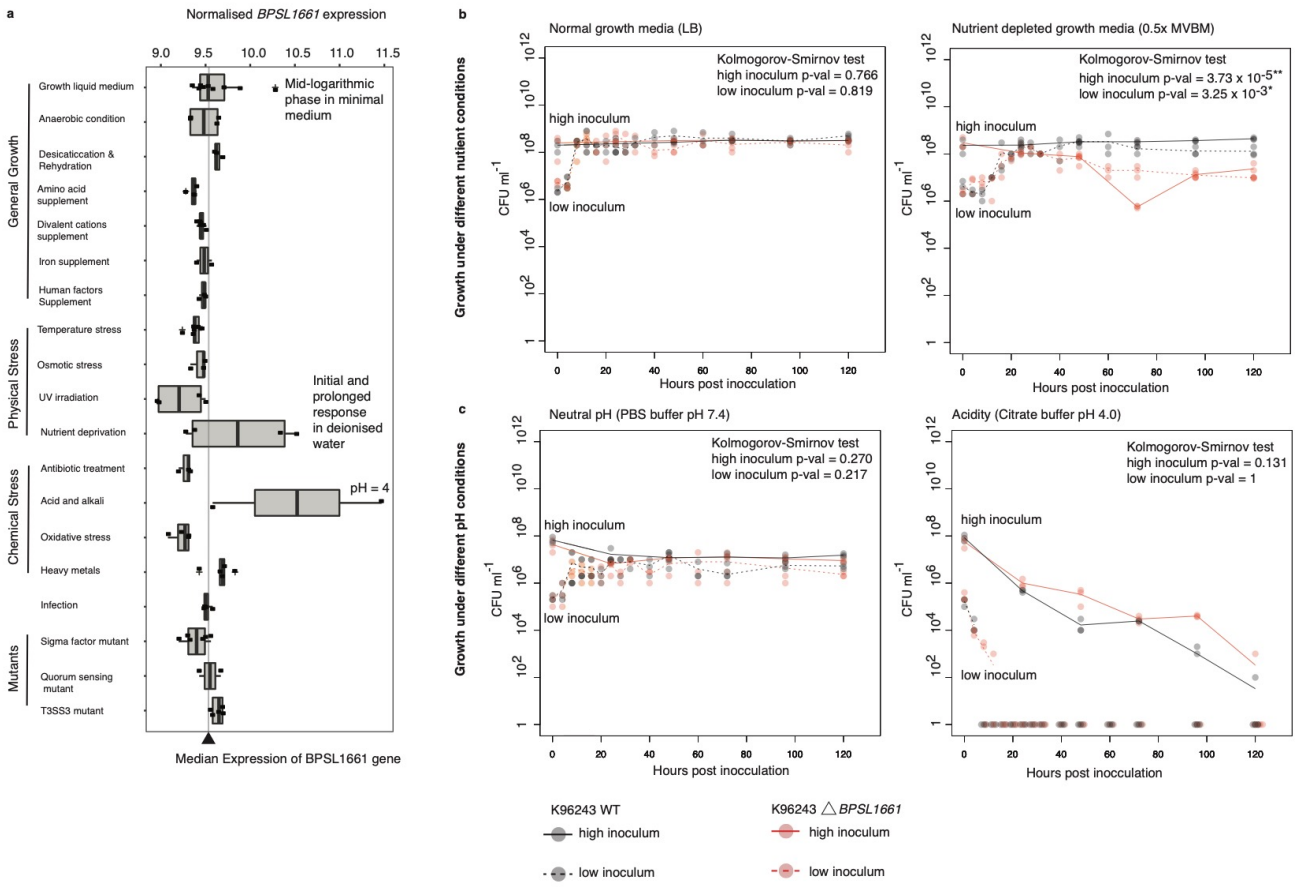


Fig 3 Functional characterisation of *BPSL1661.* a) Expression profile of *BPSL1661* across different conditions from Ooi *et al*. 2013, highlighting the gene upregulation during nutrient limited condition and high acidity. b) & c) represent growth- and stationary-phase survival of wild-type and *BPSL1661* knockout mutant from K96243 reference strain under changes in nutrition (b) and pH (c). Bacterial survival post inoculation at low (10^6 CFU ml^{-1}) and high (10^8 CFU ml^{-1}) concentration was determined as bacterial colony forming unit (CFU) at different time intervals. Three replicates were taken at each timepoint. Black and red dots denote observations from wild-type and *BPSL1661*mutant, respectively. Dotted and solid line represent profile of low- and high bacterial inoculum, respectively. Difference in growth profiles between wild-type and mutant was measured using two-sided Kolmogorov-Smirnov test.

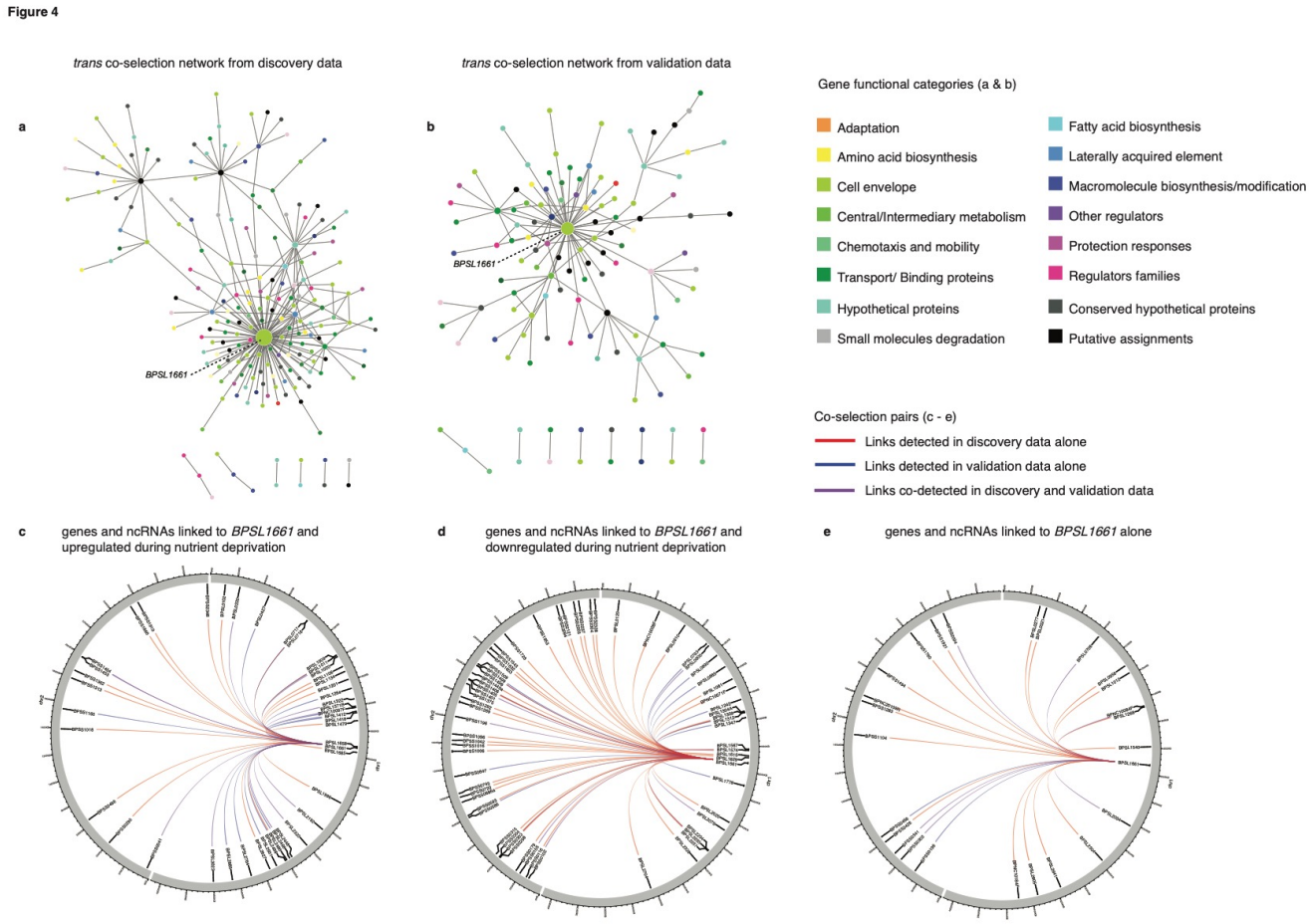
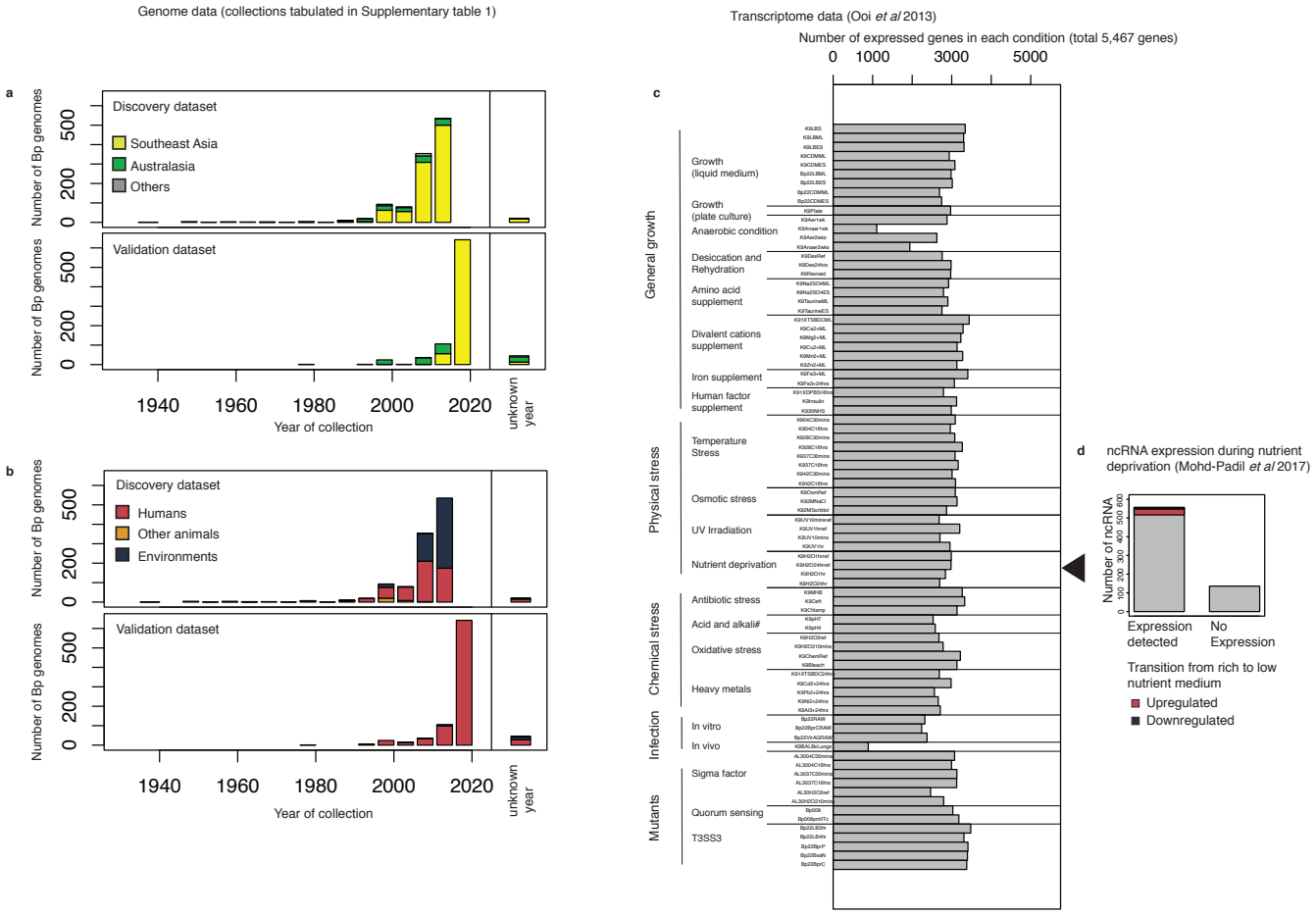


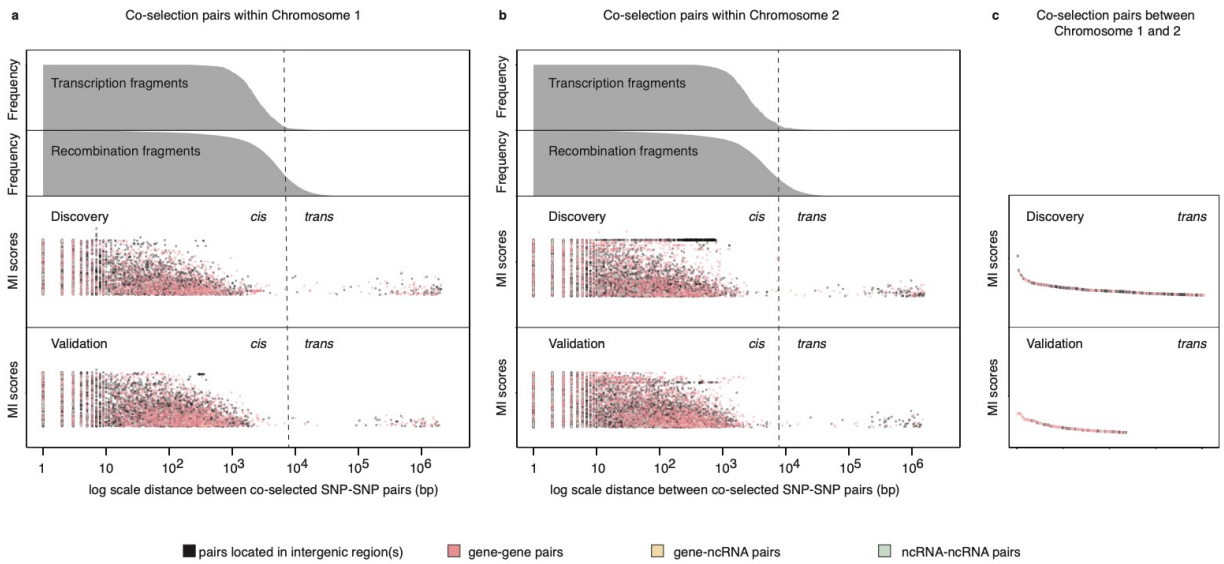
Fig 4 *BPSL1661* co-selected gene and ncRNA network. a) & b) represent networks of *trans* interacting co-selected gene-gene pairs in the discovery and validation datasets, respectively. Each node denotes a gene under co-selection with the node size proportional to numbers of pairs linked to the gene, and colour coded by the gene functional category. For both discovery and validation datasets, *BPSL1661* consistently acts as hub of the *trans* co-selected gene network. c-e) summarise genes and ncRNA co-selected with *BPSL1661* in both discovery and validation dataset: genes and ncRNA co-selected and upregulated under nutrient deprivation (c); genes and ncRNA co-selected but downregulated under nutrient deprivation (d); and genes and ncRNA that are co-selected alone (e). Links identified from the discovery data alone, validation data alone, and both data are coloured as red, blue, and purple, respectively.

Extended Data Figure 1

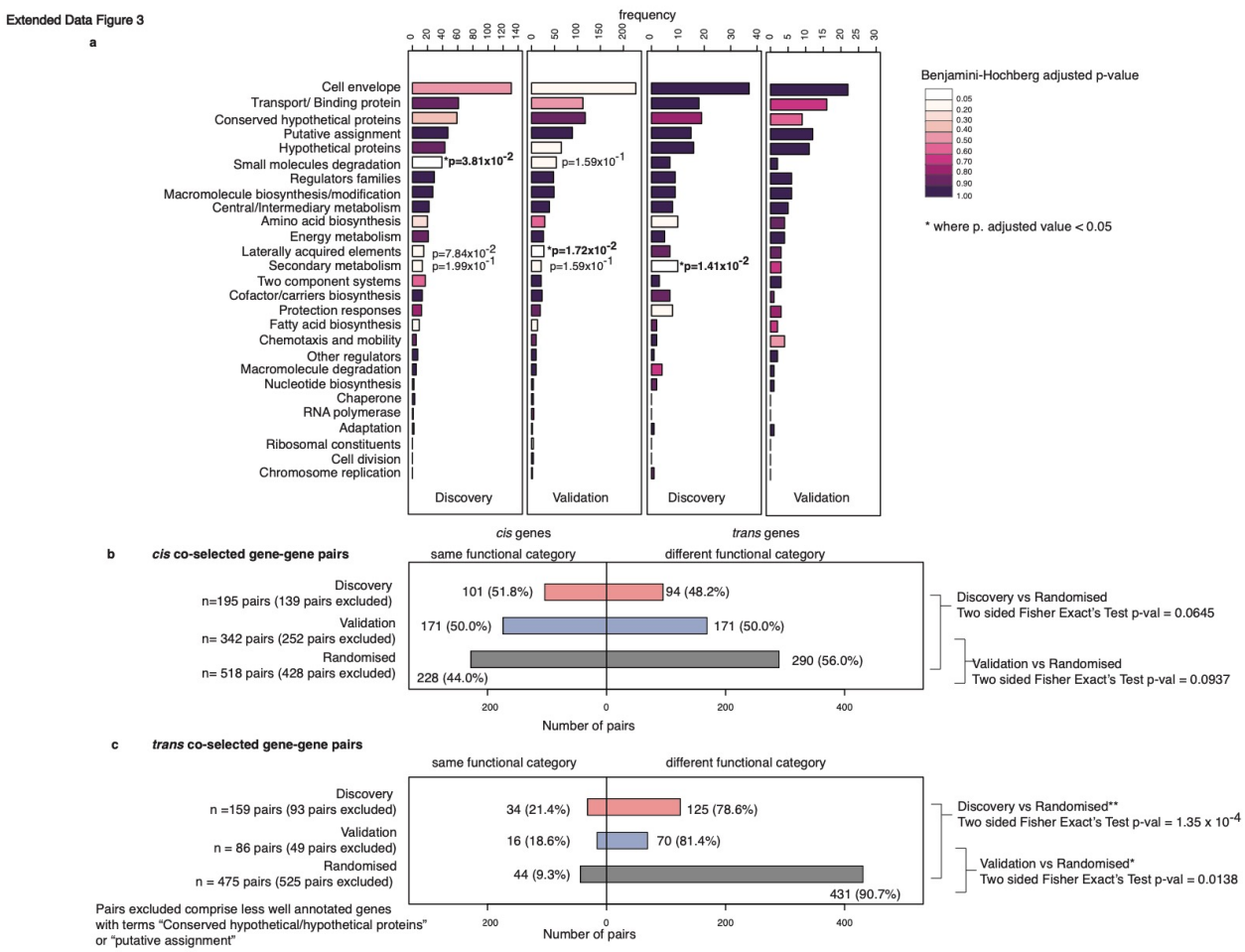


Extended Data Fig 1 Summary of genomic and transcriptomic data used for the analyses. a) - b) represent distribution of *B. pseudomallei* genomes in the discovery and validation dataset by year of collection with geographical origins and source of isolates plotted in a) and b) respectively. c) summarises *B. pseudomallei* expression profile across 82 conditions covering general growth, physical stress, chemical stress, infection and mutants obtained from from Ooi *et al.* [14] The bar chart highlights the number of genes expressed in each condition. d) represents the expression profile of *B. pseudomallei* non-coding RNA in nutrient-rich and nutrient-limited conditions. The data was obtained from Modh-Padil *et al*.[18]

Extended Data Figure 2

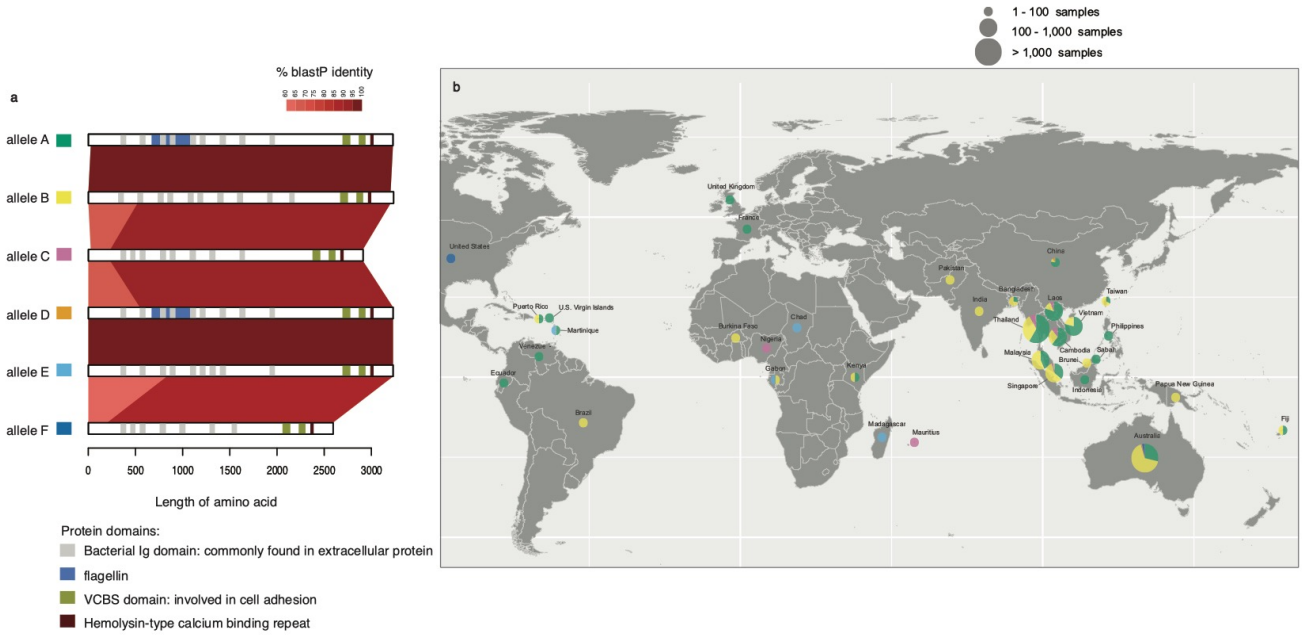


Extended Data Fig 2 Co-selected SNP-SNP pairs. a) - c) represent co-selection sites detected on Chromosome 1, Chromosome 2, and between Chromosome 1 and 2, respectively. For a) & b) The horizontal axis denotes physical distance on the chromosome on the log10 scale. The top to bottom panels represent the accumulated size of transcription fragments, the accumulated size of recombination fragments, the distance between co-selected SNP-SNP pairs against the mutual information scores from the discovery datasets, and the validation datasets, respectively. For two bottom panels, each dot represents a co-selected SNP-SNP pair and is colour-coded by their associations with genes (pink), a gene and a molecule of non-coding RNA (orange), molecules of non-coding RNA (green), or other intergenic regions (black). Vertical dotted lines denote 95th percentile of transcription fragments. c) Top to bottom panel represent the mutual information scores of each co-selected SNP-SNP pair identified in the discovery and validation datasets, respectively. The same colour scheme was employed as in a) & b).

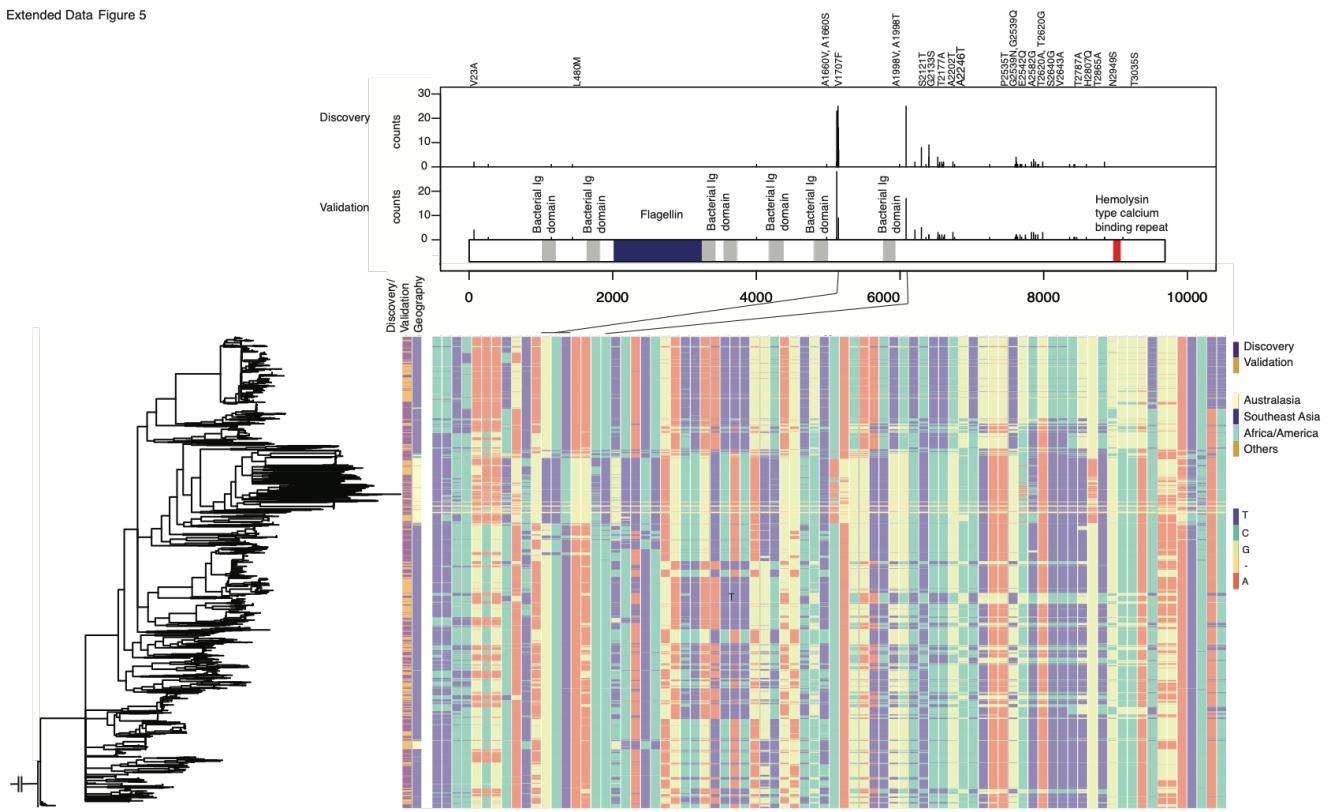


Extended Data Fig 3 Functional category of co-selected genes. (a) Enrichment of functional category observed for *cis*- and *trans*- interactions in both discovery and validation datasets. The colour indicates the significance of the enrichment test result, ranging from white (adjusted p-value <0.05) to dark purple (adjusted p-value = 1). b) & c) represent distribution of co-selected gene-gene pairs with same (left) and different (right) functional categories observed for *cis*- (b) and *trans*- (c) interactions. Data obtained from the discovery, validation, randomised dataset were highlighted as red, blue, and grey, respectively. The distribution of pairs with same and different functional categories were compared against the randomised dataset using two-sided Fisher's Exact Test. Pairs with ambiguous annotations comprising terms "conserved hypothetical/hypothetical proteins" or "putative assignment" were excluded from the analyses.

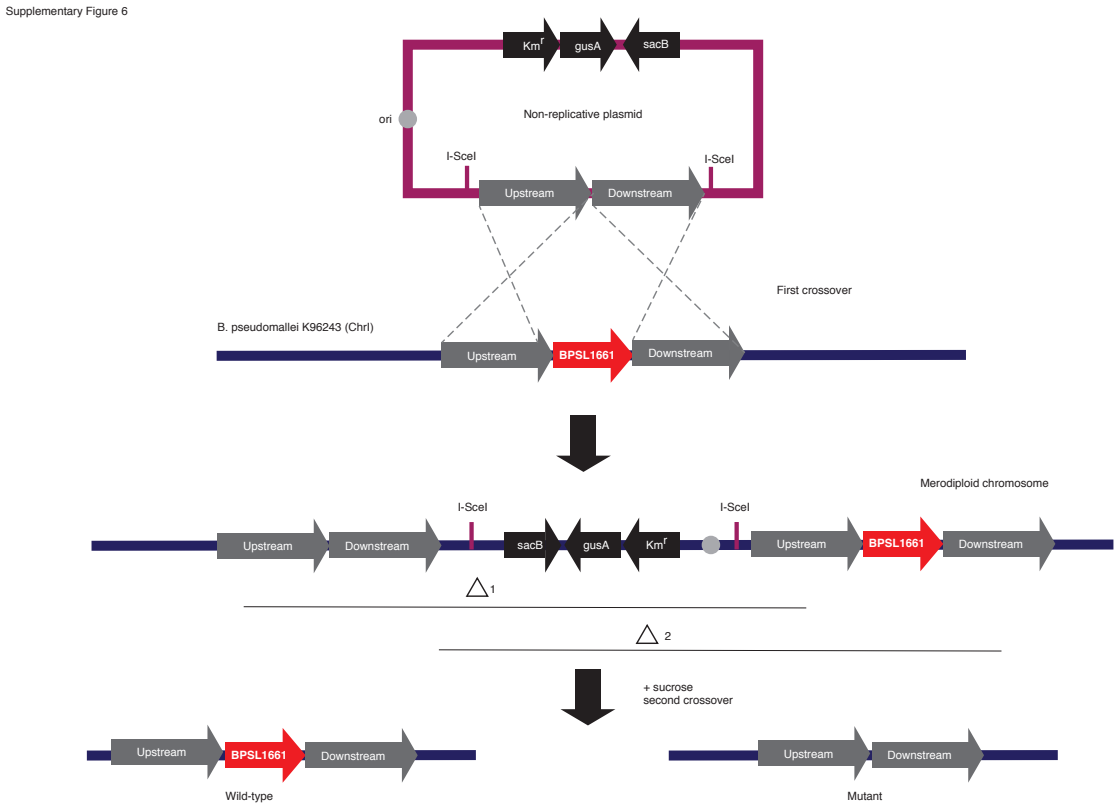
Extended Data Figure 4



Extended Data Fig 4 Distinct *BPSL1661* alleles and their geographical distribution. a) represents diagram of six different *BPSL1661* alleles (denoted A to F). Each allele harbours different combination of protein domains including bacterial Ig domain (grey); flagellin (blue); Vibrio, Colwellia, Bradyrhizobium, and Shewanella domain (VCBS, green); and hemolysin-type calcium binding repeat (brown). Red areas show protein identity. b) a world map shows geographical distribution of different *BPSL1661* alleles. Colour in pie charts correspond to different BPSL1661 alleles, with the size of the pie charts proportional to number of samples obtained from each area.



Extended Data Fig 5 Distribution of *BPSL1661* variants. a) highlights SNP locations on *BPSL1661* where co-evolutionary signals were detected in discovery (top) and validation datasets (bottom). Peaks corresponding to nonsynonymous substitutions were labelled. b) represents distribution of alleles by *B. pseudomallei* core-genome phylogeny. The estimated phylogeny is shown on the left with the first two columns labelled by type of data (discovery or validation dataset) and geographical origins of isolates, respectively. The remaining columns demonstrate all nucleotide variants detected in *BPSL1661*. Variants coincide with peaks of co-evolutionary signals were marked, one of which correspond to V1707F substitution.



Extended Data Fig 6 Construction of *BPSL1661* mutant. Schematic diagrams of plasmid-based allelic exchange of *B. pseudomallei* K96243 to construct *BPSL1661* mutant as modified from Lopez *et al.* 2009[49]. (a) a crossed PCR amplicon was generated from the assembled segments containing the upstream and downstream sequence of *BPSL1661* and cloned into pEXKM5, to construct the recombinant pEXKM5Δ*BPSL1661*;Km^r. (b) The recombinant plasmids were integrated into chromosome of *B. pseudomallei* by homologous recombination (dash line denoted as the 1st crossover). The integrated recombinant plasmid replaced the *BPSL1661* in the *B. pseudomallei* K96243 chromosome, called merodiploid. The sucrose counter selection was used to select for double crossovers (2nd crossover), resulting in generation of either in a wild-type strain (Δ1) or *BPSL1661* mutant strain (Δ2).

Extended Data table 1 Summary of bacterial strains, plasmids and oligo used in this study.

Supplementary table 1 Isolate and accession codes for both short reads and assemblies used in the discovery and validation cohorts.

Supplementary table 2 A summary of co-selected SNP-SNP sites and their mutual information scores detected in both discovery and validation datasets, in discovery dataset alone, or in validation datasets alone. The coordinates of the SNPs are based on *B. pseudomallei* K96243 reference.

Supplementary table 3 A summary of all co-selected gene-gene pairs, gene-ncRNA pairs and ncRNA-ncRNA detected in both discovery and validation datasets, in discovery dataset alone, or in validation dataset alone.

Supplementary table 4 A summary of genes (spreadsheet 1) and non-coding RNA (spreadsheet 2) co-selected with *BPSL1661* and their expression profile under different nutrient conditions.

Materials and methods

Whole-genome sequencing collections, mapping and annotation

The global *B. pseudomallei* collections[12,13,15,27,50–65] and newly sequenced data used in this study are divided into a discovery dataset (1,136 isolates) and a validation dataset (875 isolates) (**Extended Data Fig 1a & b)**. Their accession numbers are tabulated in **Supplementary table 1**. The two datasets overlap geographically and temporally, spanning two major melioidosis endemic regions of Southeast Asia and Australia. These isolates came from environmental, animal and human sources with the latter constituted the larger proportion due to availability of microbiology laboratories embedded in the clinical settings. However, over 91% of human cases represent recent acquisition from the environmental sources[66–68], thereby reducing the chance of co-evolutionary signals being shaped by human infection. For newly sequenced data, DNA libraries were sequenced on an Illumina Hiseq2000 with 100-cycle paired-end runs. Short reads were mapped to the reference *Burkholderia pseudomallei* genome K96243 using SMALT 0.7.6 [69]. Fully sequenced chromosomes and long reads sequenced by other platforms were shredded to create 100 bp paired-end reads before mapping. The genome K96243 was re-annotated using curated genes, non-coding RNA (ncRNA), and functional predictions obtained from experiments described in[14,18] (**Extended Data Fig 1c & d**).

Genome-wide epistasis and co-selection analysis

Co-selection analysis was performed on the sequence alignment using the mutual information based GWES tool SpydrPick[16] (https://github.com/santeripuranen/SpydrPick). For discovery dataset (n = 1,136), the alignment contained 389,476 SNPs, of which 206,019 and 183,457 were located in chromosome I and II, respectively. For validation dataset (n = 875), the alignment contained 285,543 SNPs, of which 150,499 and 135,044 SNPs are distributed on chromosome 1 and 2 respectively. Polymorphic sites with minor allele frequency greater than 1% and gap frequency smaller than 15% were included in the analysis. To adjust for the population structure, sequence reweighting was applied using the default similarity threshold of 0.10. An LD threshold of 7.68 kb (representing 95th percentile of transcription fragments) was used to define the *cis* and *trans* relationship between co-selected loci. Direct links for which the mutual information exceeded the extreme outlier threshold, also after removing the influence of gaps, were selected for further examination.

Functional classification of genes under co-selection

SNP-SNP pairs identified in GWES were mapped by their coordinates to genes and ncRNA annotated in K96243 reference genome[14], resulting in co-selection of gene-gene, gene-ncRNA, and ncRNA-ncRNA pairs. SNP-SNP pairs located within the same gene, or

ncRNA were removed for this analysis. Different gene functional categories including a curated Riley's classification[14], clusters of orthologous groups (COGs) [14], KEGG Pathway [70], and gene ontology (GO terms)[71] were assigned to each gene-gene pair. The final analysis was focused on a curated Riley's classification as it covered more genes detected for co-selection than other classification systems. For both discovery and validation datasets, we searched for enrichment of gene functions among co-selected genes against their distribution in K96243 genome using two-sided Fisher's Exact test (R function fisher.test()) while controlling for false positive from multiple testing using Benjamini-Hochberg method[72].

We next tested for functional dependency of co-selected gene-gene pairs by comparing the frequency of pairs with the same functional annotations to pairs with different functions. To test whether this was observed by chance, we simulated a set of 1,000 random gene-gene pairs from the K96243 genome with co-selected excluded. A separate randomised collection was pooled for *cis* (genes located < 7.68 kb apart) and *trans* interactions (genes located > 7.68 kb apart). Two-sided Fisher's Exact test was separately performed to compare the distribution of co-selected pairs with the same and different gene functions against randomised datasets from both discovery and validation datasets. Given that the bacterial genes are organised into operons, the test was also separately conducted for *cis*- and *trans* interactions where paired genes are likely under the same and different operons, respectively.

Expression patterns of genes and ncRNA under co-selection

We used *B. pseudomallei* condition-specific expression comprising 165 array profiles generated from Ooi *et al*[14] to elucidate the function of co-selected gene-gene pairs. The data had been log-transformed to fit a Gaussian distribution[14]. Expression correlations between gene-gene pairs were defined by Pearson correlation using R function cor.test(gene1, gene2, method = "pearson"). We applied Benjamini-Hochberg adjusted correction for multiple testing. Expression profiles were categorised into 5 major categories spanning general growth, physical stresses, chemical stresses, infection and mutant conditions. For all conditions, and each of five major expression condition; we compared significance correlation of expression between co-selected gene-gene pairs and randomised gene-gene pairs using a nonparametric two-sided Mann-Whitney U test. R function wilcox.test() was employed. Individual test was performed for both discovery and validation datasets, and for both *cis*- and *trans* category. Expression profile of *BPSL1661* was obtained from Ooi *et al* [14] and cross-checked for consistency with [73–75] where conditions were overlapped.

Additional pattern of ncRNA expression under nutrient limitation was sought from Mohd-Padil *et al* [18]. The authors measured and compared ncRNA expressed when *B. pseudomallei* was subjected to nutrient rich BHIB media and nutrient-depleted M9 condition.

Genetic variations in *BPSL1661* and their distribution across the core genome phylogeny

Where complete genomes were available, BLAT v. 36 [76] was used to locate the position of *BPSL1661* homolog and further confirmed with genome annotations. Illumina sequenced short reads were assembled as in [15] and annotated using Prokka v.1.14.5 [77]. Coding sequence of *BPSL1661* was identified using BLAT and confirmed with gene annotation. *BPSL1661* sequences from all genomes were aligned using MAFFT v.7.407 [78] and assigned into different clusters using CD-HIT-EST v.4.8.1 [79] with sequence identity threshold of 0.9. Here, any clusters with ≥ 5 members were considered representative alleles, resulted in 6 *BPSL1661* alleles across the datasets. Protein domains of each *BPSL1661* allele were sought from CDD/SPARCLE v.3.17 conserved domain database [80].

To investigate the distribution of *BPSL1661* variants across *B. pseudomallei* population, core-genome phylogeny was constructed from core genome SNP alignment. SNPs were called from K96243 mapped genome alignment using SNP-sites v.2.5.1 [81] excluding sites associated with mobile genetic elements [30]. We next estimated a maximum likelihood tree using IQ-TREE v.1.6.10 [82] using General Time Reversible (GTR) + Gamma distribution model of nucleotide substitution with default heuristic search options and 1,000 bootstraps[83].

***BPSL1661* functional characterisation**

The hub of co-selection signals, *BPSL1661* was next functionally characterised. Bacterial strains, plasmids and oligonucleotides used in this study are listed in **Extended Data Table 1**. GF-1 bacterial gDNA extraction kit and deoxynucleotide triphosphates (dNTPs) were purchased from Vivantis; Platinum DNA Taq polymerase from Invitrogen; pGEM-T Easy vector systems from Promega; KOD Plus DNA polymerase from Toyobo; Restriction enzymes from New England BioLabs; QIAquick Gel Extraction kit, MinElute PCR purification kit from Qiagen; Ampicillin (Ap), Kanamycin (Km), Gentamycin (Gm) were purchased from Sigma, and Isopropyl β-D-1-thiogalactopyranoside (IPTG), 5-bromo-4-chloro-3-indolyl-β-D-glactopyranoside (X-Gal) and 5-bromo-4-chloro-3-indolyl-β-D-glucuronide (X-Gluc) were purchased from Gold Biotechnology.

The culture of *B. pseudomallei* K96243 wild-type, *B. pseudomallei BPSL1661* clean mutant, and the *Escherichia coli* strains used for construction of the *B. pseudomallei* mutant were routinely grown in Luria–Bertani (LB) and Luria-Lennox (LB, low-salt) medium (Sigma,

USA), at 37°C with 200 rpm agitation. When necessary, the medium was supplemented with antibiotics, chemicals and chromogens at the concentrations of 100 µg ml^{-1} Ap, 35 µg ml^{-1} Km, 0.1 mM IPTG, 50 µg ml^{-1} X-Gal for *E. coli*; 5 µg ml^{-1} Gm, 1,000 µg ml^{-1} Km, 0.1 mM IPTG, 50 µg ml^{-1} X-Gluc for *B. pseudomallei.*

Construction of the *BPSL1661* clean deletion mutant

The nucleotide sequence encoding (*BPS_RS08795*, [*BPSL1661*]) gene (GenBank accession no. WP_045606470.1) of *B. pseudomallei* K96243 was used to design primers for clean deletion using Primer-BLAST program [84]. The BPSL1661 clean mutant (Δ*BPSL1661*) (**Extended Data Fig 6**) was constructed from *B. pseudomallei* K96243 by double-crossover allelic exchange as described previously[49]. Briefly, the upstream and downstream DNA fragments of *BPSL1661* were amplified using BPSL1661_PFup/PRup and BPSL1661_PFdown/PRdown primers and then ligated. The ligated amplicons were cloned into TA cloning vector pGEM-T Easy (Promega, USA) and check for its correct insert size before subcloned into an allelic exchange plasmid pEXKM5 [49]. The recombinant pEXKM5 plasmids were introduced into *B. pseudomallei* K96243 by bi-parental conjugation using conjugal *E. coli* S17-llpir[85]. Merodiploid was selected on LB medium supplemented with X-Gluc and Kanamycin that appeared as pale blue colonies. To obtain the clean deletion Δ*BPSL1661* mutant, merodiploid colonies were cultured in LB medium to reach stationary phase then subculture into YT broth supplemented with 10 % sucrose for overnight and spread on LB-medium with 15% sucrose. Suspected *B. pseudomallei* colonies was verified by PCR and confirmed by DNA sequencing of the region flanking *BPSL1661*.

***BPSL1661* condition-specific assays**

Guided by expression profile of *BPSL1661* in Ooi *et al*[14], the growth of wild-type and Δ*BPSL1661* were compared by enumerating the bacterium grown under normal growth (LB broth), nutrient depleted condition (MVBM), neutral pH and acidic pH. The bacterium was grown on Ashdown's selective agar[86] before transferring to grow under each condition.

We followed Ooi *et al*[14] for normal growth condition using LB broth. We artificially induced nutrient limited condition *in vitro* by replacing glucose limited media (six carbon source) with glycerol (three carbon source) in MVBM (Modified Vogel and Bonner's medium). Morever, MVBM [87] further forced cell starvation by inducing biofilm formation. Nutrients will be consumed by cells located on the periphery of the biofilm clusters leading to reduced level of nutrients diffused to the inner cells. Both wild-type and Δ*BPSL1661* were grown on 0.05% glucose in 0.5 x MVBM [0.05%G, 0.5×MVBM]) at 37°C with 200 rpm shaking for overnight. The culture was adjusted and measured an optical density (OD) at

OD600 nm to be 0.2 (OD_{600} = 0.2) and then inoculated into 10 ml prewarmed 0.05%G, 0.5×MVBM medium at 37°C with 100 rpm shaking.

Growth in neutral pH was conducted in 10-fold diluted phosphate buffered saline (PBS) pH 7.4. Growth in a low pH was done as described previously [3]. For both condition, bacteria were cultured for overnight in LB broth at 37°C with 200 rpm shaking and sub-cultured in new fresh medium until log phase. Bacterial cells were washed in 1×PBS buffer pH 7.4 or 1×citrate buffer pH 4.0 for three times by centrifugation (8000 × g for 5 min) and resuspended in 10 ml of each buffer.

For all conditions, we observed bacterial initial growth and stationary-phase survival for both low (10^6 CFU ml^{-1}) and high (10^8 CFU ml^{-1}) inoculum. The colony forming unit (CFU) post inoculation was counted at different time interval. All assays involved three replicates that were independently prepared, cultured and treated. Difference in growth profile between wild-type and Δ*BPSL1661* for each condition was compared by a nonparametric Kolmogorov-Smirnov test[88].

Visualisation

Visualisation of phylogenetic trees was performed in Phandango[89]. Figures were plotted using R[90] and R applications[91].

Statistics and reproducibility

No sample sizes were predetermined. For categorical data, two-sided Fisher's exact tests were employed to compare differences between each group. For continuous data, non-parametric Mann-Whitney U tests and a parametric Pearson correlation were used to compare the inter-group distribution and the correlation of the expression levels between each co-selected gene-gene pair, respectively. Where appropriate, we used Benjamini-Hochberg procedure to adjust p value for multiple comparisons.

Data availability

Short reads for isolates are archived in ENA or NCBI database. Accession number for each individual isolate in discovery and validation dataset are provided in Supplementary Table 1. Raw results of co-evolutionary mutation pairs, co-selected gene-gene pairs, and *BPSL1661* co-selection gene and ncRNA network are given in Supplementary Table 2, 3, and 4, respectively.

Code availability

The code used to detect co-evolutionary signals is available in https://github.com/santeripuranen/SpydrPick

Acknowledgements

The authors thank the Wellcome Sanger Institute Library construction, sequence and core informatics teams, the pathogen informatics team for their technical support. CC was funded by Wellcome International Intermediate Fellowship (216457/Z/19/Z), Sanger International Fellowship and Thailand Research Fund (MRG6280226). JC was funded by the ERC grant no. 742158. 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), and to the Wellcome Thailand and African Programme (106698).

Author contributions

J.C. conceived the original concept. C.C., R.W.S. and J.C. designed the experiment and oversaw the project and analyses. C.C., S.J.P., N.P.J.D., J.Parkhill, N.R.T, R.W.S, and J.C. acquired funding. P.B., K.S., E.J.F., S.D., N.C., D.L., and N.P.J.D. contributed samples and reagents or assisted in samples preparation and permitting. J.Pensar and M.P. developed analytical tools. C.C. and J.Pensar performed genomic analyses. S.C. developed gene knockout mutant. S.C. and C.P. performed experimental validations. C.C., J. Pensar, A.S., E.J.F., N.C., D.L., J.Parkhill, N.R.T, R.W.S. and J.C. contributed to the interpretation and presentation of results in the main manuscript and supplementary documents. C.C. and J.C. wrote the first draft with input from all authors.

Competing financial interests

The authors declare no competing financial interests.

References (ref 1 – 48 are included in the main text)

1. Limmathurotsakul, D. *et al.* Predicted global distribution of *Burkholderia pseudomallei* and burden of melioidosis. *Nat. Microbiol.* **1**, 15008 (2016).
2. Wiersinga, W. J. *et al.* Melioidosis. *Nat. Rev. Dis. Primer* **4**, 17107 (2018).
3. Inglis, T. J. J. & Sagripanti, J.-L. Environmental Factors That Affect the Survival and Persistence of *Burkholderia pseudomallei*. *Appl. Environ. Microbiol.* **72**, 6865–6875 (2006).
4. Limmathurotsakul, D. *et al.* Systematic Review and Consensus Guidelines for Environmental Sampling of *Burkholderia pseudomallei*. *PLoS Negl. Trop. Dis.* **7**, e2105 (2013).
5. Hantrakun, V. *et al.* Soil Nutrient Depletion Is Associated with the Presence of *Burkholderia pseudomallei*. *Appl. Environ. Microbiol.* **82**, 7086–7092 (2016).
6. Baker, A. L., Ezzahir, J., Gardiner, C., Shipton, W. & Warner, J. M. Environmental Attributes Influencing the Distribution of *Burkholderia pseudomallei* in Northern Australia. *PLOS ONE* **10**, e0138953 (2015).
7. Manivanh, L. *et al. Burkholderia pseudomallei* in a lowland rice paddy: seasonal changes and influence of soil depth and physico-chemical properties. *Sci. Rep.* **7**, 3031 (2017).
8. Ngamsang, R. *et al.* THE CONTRIBUTION OF SOIL PHYSICOCHEMICAL PROPERTIES TO THE PRESENCE AND GENETIC DIVERSITY OF *BURKHOLDERIA PSEUDOMALLEI*. *Southeast Asian J. Trop. Med. Public Health* **46**, 38–50 (2015).
9. Coenye, T. & Vandamme, P. Diversity and significance of *Burkholderia* species occupying diverse ecological niches. *Environ. Microbiol.* **5**, 719–729 (2003).
10. Yip, T.-W. *et al.* Endemic Melioidosis in Residents of Desert Region after Atypically Intense Rainfall in Central Australia, 2011. *Emerg. Infect. Dis.* **21**, 1038–1040 (2015).
11. Pumpuang, A. *et al.* Survival of *Burkholderia pseudomallei* in distilled water for 16 years. *Trans. R. Soc. Trop. Med. Hyg.* **105**, 598–600 (2011).

12. Chewapreecha, C. *et al.* Global and regional dissemination and evolution of *Burkholderia pseudomallei*. *Nat. Microbiol.* **2**, 16263 (2017).
13. Spring-Pearson, S. M. *et al.* Pangenome Analysis of *Burkholderia pseudomallei*: Genome Evolution Preserves Gene Order despite High Recombination Rates. *PLOS ONE* **10**, e0140274 (2015).
14. Ooi, W. F. *et al.* The Condition-Dependent Transcriptional Landscape of *Burkholderia pseudomallei*. *PLoS Genet.* **9**, e1003795 (2013).
15. Chewapreecha, C. *et al.* Genetic variation associated with infection and the environment in the accidental pathogen *Burkholderia pseudomallei*. *Commun. Biol.* **2**, 428 (2019).
16. Pensar, J. *et al.* Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res.* **47**, e112–e112 (2019).
17. Schubert, B., Maddamsetti, R., Nyman, J., Farhat, M. R. & Marks, D. S. Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nat. Microbiol.* **4**, 328–338 (2019).
18. Mohd-Padil, H. *et al.* Identification of sRNA mediated responses to nutrient depletion in *Burkholderia pseudomallei*. *Sci. Rep.* **7**, 17173 (2017).
19. Lagator, M., Paixão, T., Barton, N. H., Bollback, J. P. & Guet, C. C. On the mechanistic nature of epistasis in a canonical cis-regulatory element. *eLife* **6**, e25192 (2017).
20. Guil, S. & Esteller, M. Cis-acting noncoding RNAs: friends and foes. *Nat. Struct. Mol. Biol.* **19**, 1068–1075 (2012).
21. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
22. Lemos, B., Araripe, L. O., Fontanillas, P. & Hartl, D. L. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc. Natl. Acad. Sci.* **105**, 14471–14476 (2008).
23. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Independent Effects of *cis* - and *trans* - regulatory Variation on Gene Expression in *Drosophila melanogaster*. *Genetics* **178**, 1831–1835 (2008).

24. Metzger, B. P. H. *et al.* Contrasting Frequencies and Effects of *cis* - and *trans* - Regulatory Mutations Affecting Gene Expression. *Mol. Biol. Evol.* **33**, 1131–1146 (2016).
25. Winsor, G. L. *et al.* The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics* **24**, 2803–2804 (2008).
26. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Rep.* **14**, 679–685 (2016).
27. Holden, M. T. G. *et al.* Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci.* **101**, 14240–14245 (2004).
28. Shieh, Y.-W. *et al.* Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* **350**, 678–680 (2015).
29. 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 (2018).
30. Tuanyok, A. *et al.* Genomic islands from five strains of *Burkholderia pseudomallei*. *BMC Genomics* **9**, 566 (2008).
31. Suwannasaen, D. *et al.* Human Immune Responses to *Burkholderia pseudomallei* Characterized by Protein Microarray Analysis. *J. Infect. Dis.* **203**, 1002–1011 (2011).
32. Kohler, C. *et al.* Rapid and Sensitive Multiplex Detection of *Burkholderia pseudomallei*-Specific Antibodies in Melioidosis Patients Based on a Protein Microarray Approach. *PLoS Negl. Trop. Dis.* **10**, e0004847 (2016).
33. Felgner, P. L. *et al.* A *Burkholderia pseudomallei* protein microarray reveals serodiagnostic and cross-reactive antigens. *Proc. Natl. Acad. Sci.* **106**, 13499–13504 (2009).
34. Ribolzi, O. *et al.* Land use and soil type determine the presence of the pathogen *Burkholderia pseudomallei* in tropical rivers. *Environ. Sci. Pollut. Res.* **23**, 7828–7839 (2016).
35. Musa, H. I. *et al.* Physicochemical Properties Influencing Presence of *Burkholderia pseudomallei* in Soil from Small Ruminant Farms in Peninsular Malaysia. *PLOS ONE* **11**, e0162348 (2016).

36. Kaestli, M. *et al.* What Drives the Occurrence of the Melioidosis Bacterium *Burkholderia pseudomallei* in Domestic Gardens? *PLoS Negl. Trop. Dis.* **9**, e0003635 (2015).

37. Goodrick, I., Todd, G. & Stewart, J. Soil characteristics influencing the spatial distribution of melioidosis in Far North Queensland, Australia. *Epidemiol. Infect.* **146**, 1602–1607 (2018).

38. Müller, C. M. *et al.* Role of *RelA* and *SpoT* in *Burkholderia pseudomallei* virulence and immunity. *Infect. Immun.* **80**, 3247–3255 (2012).

39. Wang, J. D., Sanders, G. M. & Grossman, A. D. Nutritional Control of Elongation of DNA Replication by (p)ppGpp. *Cell* **128**, 865–875 (2007).

40. Wösten, M. M. S. M., van de Lest, C. H. A., van Dijk, L. & van Putten, J. P. M. Function and Regulation of the C4-Dicarboxylate Transporters in *Campylobacter jejuni*. *Front. Microbiol.* **8**, (2017).

41. Puckett, S. *et al.* Glyoxylate detoxification is an essential function of malate synthase required for carbon assimilation in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci.* **114**, E2225–E2232 (2017).

42. Wang, L. *et al.* Structure and Evolution of Glycogen Branching Enzyme N-Termini From Bacteria. *Front. Microbiol.* **9**, 3354 (2019).

43. Santín, C. & Doerr, S. H. Fire effects on soils: the human dimension. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150171 (2016).

44. Pellegrini, A. F. A. *et al.* Fire frequency drives decadal changes in soil carbon and nitrogen and ecosystem productivity. *Nature* **553**, 194–198 (2018).

45. Bowd, E. J., Banks, S. C., Strong, C. L. & Lindenmayer, D. B. Long-term impacts of wildfire and logging on forest soils. *Nat. Geosci.* **12**, 113–118 (2019).

46. Shyamsundar, P. *et al.* Fields on fire: Alternatives to crop residue burning in India. *Science* **365**, 536–538 (2019).

47. Vadrevu, K. P. *et al.* Trends in Vegetation fires in South and Southeast Asian Countries. *Sci. Rep.* **9**, 7422 (2019).

48. Kershaw, A. P. Climatic change and Aboriginal burning in north-east Australia during the last two glacial/interglacial cycles. *Nature* **322**, 47–49 (1986).

49. López, C. M., Rholl, D. A., Trunck, L. A. & Schweizer, H. P. Versatile Dual-Technology System for Markerless Allele Replacement in *Burkholderia pseudomallei*. *Appl. Environ. Microbiol.* **75**, 6496–6503 (2009).

50. Song, L. *et al.* Draft Genome Sequence of *Burkholderia pseudomallei* Strain 350105, Isolated in Hainan, China, in 1976. *Genome Announc.* **3**, (2015).

51. Daligault, H. E. *et al.* Whole-Genome Assemblies of 56 *Burkholderia* Species. *Genome Announc.* **2**, e01106-14, 2/6/e01106-14 (2014).

52. Johnson, S. L. *et al.* Whole-Genome Sequences of 80 Environmental and Clinical Isolates of *Burkholderia pseudomallei*: TABLE 1. *Genome Announc.* **3**, e01282-14, /ga/3/1/e01282-14.atom (2015).

53. Johnson, S. L. *et al.* Complete Genome Sequences for 59 *Burkholderia* Isolates, Both Pathogenic and Near Neighbor: TABLE 1. *Genome Announc.* **3**, e00159-15, /ga/3/2/e00159-15.atom (2015).

54. Price, E. P. *et al.* Unprecedented Melioidosis Cases in Northern Australia Caused by an Asian B*urkholderia pseudomallei* Strain Identified by Using Large-Scale Comparative Genomics. *Appl. Environ. Microbiol.* **82**, 954–963 (2016).

55. Sahl, J. W. *et al.* Genome Sequence of *Burkholderia pseudomallei* NCTC 13392. *Genome Announc.* **1**, (2013).

56. Viberg, L. T. *et al.* Whole-Genome Sequences of Five *Burkholderia pseudomallei* Isolates from Australian Cystic Fibrosis Patients. *Genome Announc.* **3**, (2015).

57. McRobb, E. *et al.* Tracing melioidosis back to the source: using whole-genome sequencing to investigate an outbreak originating from a contaminated domestic water supply. *J. Clin. Microbiol.* **53**, 1144–1148 (2015).

58. Nandi, T. *et al. Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res.* **25**, 129–141 (2015).

59. Hayden, H. S. *et al.* Evolution of *Burkholderia pseudomallei* in recurrent melioidosis. *PloS One* **7**, e36507 (2012).

60. Price, E. P. *et al.* Within-Host Evolution of *Burkholderia pseudomallei* over a Twelve-Year Chronic Carriage Infection. *mBio* **4**, e00388-13 (2013).

61. Chen, Y.-S. *et al.* Whole-Genome Sequence of an Epidemic Strain of *Burkholderia pseudomallei* vgh07 in Taiwan. *Genome Announc.* **3**, (2015).

62. Hsueh, P.-T. *et al.* Comparison of Whole-Genome Sequences from Two Colony Morphovars of *Burkholderia pseudomallei*. *Genome Announc.* **3**, (2015).

63. Sidjabat, H. E., Cottrell, K. & Cervin, A. Draft Genome Sequences of *Burkholderia pseudomallei* and *Staphylococcus aureus*, Isolated from a Patient with Chronic Rhinosinusitis. *Genome Announc.* **3**, (2015).

64. Bugrysheva, J. V. *et al.* Finished Annotated Genome Sequence of *Burkholderia pseudomallei* Strain Bp1651, a Multidrug-Resistant Clinical Isolate. *Genome Announc.* **3**, (2015).

65. Chapple, S. N. J. *et al.* Whole-genome sequencing of a quarter-century melioidosis outbreak in temperate Australia uncovers a region of low-prevalence endemicity. *Microb. Genomics* **2**, e000067 (2016).

66. Currie, B. J., Fisher, D. A., Anstey, N. M. & Jacups, S. P. Melioidosis: acute and chronic disease, relapse and re-activation. *Trans. R. Soc. Trop. Med. Hyg.* **94**, 301–304 (2000).

67. Wiersinga, W. J., Currie, B. J. & Peacock, S. J. Melioidosis. *N. Engl. J. Med.* **367**, 1035–1044 (2012).

68. Wiersinga, W. J. *et al.* Melioidosis. *Nat. Rev. Dis. Primer* **4**, 17107 (2018).

69. Ponstingl, H. & Ning, Z. *SMALT - A new mapper for DNA sequencing reads*. (2010).

70. Kanehisa, M. & Sato, Y. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.* **29**, 28–35 (2020).

71. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

72. Benjamini, Y. & Hochberg, Y. Controlling the false discovery Rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57**, 289–300 (1995).

73. Rodrigues, F. *et al.* Global Map of Growth-Regulated Gene Expression in *Burkholderia pseudomallei*, the Causative Agent of Melioidosis. *J. Bacteriol.* **188**, 8178–8188 (2006).

74. Jitprasutwit, S. *et al.* Identification of *Burkholderia pseudomallei* Genes Induced During Infection of Macrophages by Differential Fluorescence Induction. *Front. Microbiol.* **11**, 72 (2020).

75. Chin, C.-Y. *et al.* Global transcriptional analysis of *Burkholderia pseudomallei* high and low biofilm producers reveals insights into biofilm production and virulence. *BMC Genomics* **16**, 471 (2015).

76. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

77. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

78. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

79. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

80. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).

81. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2**, (2016).

82. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

83. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

84. Ye, J. *et al.* Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).

85. Simon, R., Priefer, U. & Pühler, A. A Broad Host Range Mobilization System for *In Vivo* Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. *Bio/Technology* **1**, 784–791 (1983).

86. Ashdown, L. R. An improved screening technique for isolation of *pseudomonas pseudomallei* from clinical specimens. *Pathology (Phila.)* **11**, 293–297 (1979).

87. Lam, J., Chan, R., Lam, K. & Costerton, J. W. Production of mucoid microcolonies by *Pseudomonas aeruginosa* within infected lungs in cystic fibrosis. *Infect. Immun.* **28**, 546–556 (1980).

88. Berger, V. W. & Zhou, Y. Kolmogorov-Smirnov Test: Overview. in *Wiley StatsRef: Statistics Reference Online* (eds. Balakrishnan, N. et al.) stat06558 (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat06558.

89. Hadfield, J. *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* **34**, 292–293 (2018).

90. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. (2017).

91. Yu, Y., Ouyang, Y. & Yao, W. shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **34**, 1229–1231 (2018).