Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen *Orientia tsutsugamushi*

Elizabeth M. Batty[a,b,c], Suwittra Chaemchuen[b], Stuart D. Blacksell[b,c], Daniel Paris[b,c,d,e], Rory Bowden[a], Caroline Chan[f], Ramkumar Lachumanan[f], Nicholas Day[b,c], Peter Donnelly[a,g], Swaine L. Chen[h,i], Jeanne Salje[b,c#]

Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX1 7BN, UK[a] ;

Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand[b] ;

Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom[c] ;

Swiss Tropical and Public Health Institute, Basel, Switzerland[d] **;**

Faculty of Medicine, University Basel, Basel, Switzerland[e] ;

Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025, USA[f] ;

Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK[g]

Department of Medicine, Division of Infectious Diseases, Yong Loo Lin School of Medicine, National University of Singapore, Singapore[h]

Genome Institute of Singapore, A*STAR, Singapore 138672[i]

[#] Address correspondence to Jeanne Salje: jeanne.salje@ndm.ox.ac.uk

Abstract (250 words)

Background

Orientia tsutsugamushi is a clinically important but neglected obligate intracellular bacterial pathogen of the Rickettsiaceae family that causes the potentially life-threatening human disease scrub typhus. In contrast to the genome reduction seen in many obligate intracellular bacteria, early genetic studies of *Orientia* have revealed one of the most repetitive bacterial genomes sequenced to date. The dramatic expansion of mobile elements has hampered efforts to generate complete genome sequences using short read sequencing methodologies, and consequently there have been few studies of the comparative genomics of this neglected species.

Results

We report new high-quality genomes of *Orientia tsutsugamushi,* generated using PacBio single molecule long read sequencing, for six strains: Karp, Kato, Gilliam, TA686, UT76 and UT176. In comparative genomics analyses of these strains together with existing reference genomes from Ikeda and Boryong strains, we identify a relatively small core genome of 657 genes, grouped into core gene 'islands' and separated by repeat regions, and use the core genes to infer the first whole-genome phylogeny of *Orientia.*

Conclusions

Complete assemblies of multiple Orientia genomes verify initial suggestions that these are remarkable organisms. They have large genomes with widespread amplification of repeat elements and massive chromosomal rearrangements between strains. At the gene level, Orientia has a relatively small set of universally conserved genes, similar to other obligate intracellular bacteria, and the relative expansion in genome size can be accounted for by gene duplication and repeat amplification. Our study demonstrates the utility of long read sequencing to investigate complex bacterial genomes and characterise genomic variation.

Keywords (3-10)

Orientia tsutsugamushi, Rickettsiales, comparative genomics, PacBio, long read sequencing

Introduction

Background

Orientia tsutsugamushi is an obligate intracellular bacterial pathogen of the order Rickettsiales, family Rickettsiaceae which causes the life-threatening human disease scrub typhus. *Orientia* is transmitted by *Leptotrombidium* mites that occasionally feed on humans during the larval stage of development ("chiggers"), inoculate bacteria into the skin, and initiate infection. *Orientia* is maintained in mite populations by transovarial transmission. The mites normally feed only once on a vertebrate host, and cannot transmit bacteria directly from one host to another (Coleman et al., 2003). Bacteria propagate within endothelial cells, dendritic cells and monocytes at the site of inoculation, sometimes resulting in a visible red skin feature called an eschar (Paris et al., 2012). Bacteria subsequently spread through the endothelial and lymphatic system to cause a systemic infection characterised by lymphadenopathy, headache, fever, rash and myalgia, which typically begin 7-10 days after inoculation. The non-specificity of these symptoms makes scrub typhus difficult to diagnose based purely on clinical observations, and this is an important reason why the prevalence of scrub typhus has been historically under-recognised. Scrub typhus has now been shown to be a leading cause of severe fever and sepsis in studies in Thailand, India, China, Laos and Myanmar (Luce-Fedrow et al., 2018) and untreated or severe cases are associated with CNS infection, morbidity and death (Bonell et al., 2017; Dittrich et al., 2015). Locally acquired cases of scrub typhus have been reported in South America and the Middle East(Izzard et al., 2010; Weitzel et al., 2016), suggesting that the global burden of this disease may stretch beyond the traditionally known endemic areas of Asia and Northern Australia (Luce-Fedrow et al., 2018).

Orientia tsutsugamushi (previously *Rickettsia tsutsugamushi),* is distinct from other members of the Rickettsiaceae. The genus Orientia currently includes two known species, *O. tsutsugamushi* and *O. chuto,* the latter represented to date by a single strain isolated from a patient with a febrile illness contracted in Dubai (Izzard et al., 2010). High antigenic diversity among strains of *Orientia tsutsugamushi* is reflected in the poor immunological protection that recovered patients exhibit towards strains different from their original infection and, combined with a complex immune response that involves both humoral and cell-mediated immunity, this has hampered efforts towards vaccine development.

Despite its importance as a pathogen, few genomic analyses of *O. tsutsugamushi* have been published. The first whole genome sequence, Boryong, (Cho et al., 2007) reported a proliferation of type IV secretion systems in a repeat-dense genome of which 37.1% comprised identical repeats. A comparison of Boryong and the second complete genome, Ikeda (Nakayama et al., 2008), revealed similar repeats present in each genome, dominated by an integrative element named the *Orientia tsutsugamushi* amplified genetic element (OTage), and identified a core genome of 520 genes shared between the two *O. tsutsugamushi* strains and the 5 available sequences of other *Rickettsia* (Nakayama et al., 2010). Extensive genomic reshuffling was thought to have been mediated by amplification of repetitive sequences.

In comparison to other *Rickettsiae*, many of which have small and extremely stable genomes, *Orientia tsutsugamushi* has a large genome with an extraordinary proliferation of repeat sequences and conjugative elements. Some of the conjugative elements present in multiple copies across the genome are homologues of a gene cluster found in a single copy in *Rickettsia bellii*. Many of the genes in these elements are fragmented, suggesting they are non-functional (Darby et al., 2007). Other intracellular pathogens also contain repetitive elements, such as the mobile genetic elements in *Wolbachia* (Wu et al., 2004) and the tandem intergenic repeats in *Ehrlichia ruminantum* (Frutos et al., 2006). These mechanisms may evolve to increase genetic variability and aid immune evasion in bacteria which cannot easily take up novel DNA.

Larger collections of *O. tsutsugamushi* strains have been extensively studied using MLST and sequence typing of the *groES* and *groEL* (Arai et al., 2013) genes, and the outer membrane proteins 47kDa (also called HtrA or TSA47) (Jiang et al., 2013) and 56kDa (also called OmpA or TSA56) (Lu et al., 2010) genes. The 56kDa and 47kDa genes are highly immunogenic in human patients and animal models and have long been investigated as candidates for vaccine design, but high levels of diversity between strains, especially in the 56kDa gene, have limited the potential of developing a universal vaccine based on these epitopes.

Multiple studies in South East Asia have looked at the diversity of strains by MLST and 56kDa typing, and shown a high level of diversity, with many MLSTs unique to an individual strain (Duong et al., 2013; Phetsouvanh et al., 2015; Sonthayanon et al., 2010; Wongprompitak et al., 2015). Work in Thailand and Laos has shown recombination between MLSTs, as well as evidence for multiple infections in individual patients, implying that multiple strains may co-exist in mites (Sonthayanon et al., 2010). Comparisons of *56kDa* typing with MLST (Sonthayanon et al., 2010) and *47kDa* (Jiang et al., 2013) also show low congruence between methods, suggesting that single gene typing of *Orientia* may not represent the true relationships between strains; by extension, a 7-gene MLST scheme may not capture the full set of genomic relationships among strains.

Attempts to generate complete *Orientia tsutsugamushi* genomes by whole genome sequencing have been limited by the difficulties of sequencing and assembling a repeat-dense genome, and no further genomes have been completed since the Boryong and Ikeda genomes in 2008. Current draft assemblies are fragmented with over 50 contigs per genome, and vary in size – the two assemblies of the genome of *Orientia tsutsugamushi* str. Karp available on Genbank are 1,459,958bp (https://www.ebi.ac.uk/ena/data/view/LANM01000000) and 2,022,909bp (https://www.ncbi.nlm.nih.gov/nuccore/LYMA00000000; Liao et al., 2017) in length, suggesting that assemblies are either incomplete, or have problems caused by the misassembly of repeats or the inclusion of contaminating sequences.

In this work, we have used Pacific Biosciences long-read sequencing to assemble six complete genomes of *Orientia tsutsgamushi* strains representing a range of geographical origins and serotypes. From this, we gain new insights into potential mechanisms underlying the characteristic antigenic diversity of Orientia, which may contribute to its widespread prevalence among humans. Finally, this expanded genomic perspective will contribute to our understanding of the phylogeography and epidemiology of this species, as well as contribute to more detailed studies of virulence mechanisms.

Methods

Bacterial propagation

All experiments were performed using *O. tsutsugamushi* grown in the mouse fibroblast cell line L929. Uninfected L929 cells were grown in 25 cm^2 and 75 cm^2 plastic flasks at 37 ^oC and 5% CO_2, using DMEM or RPMI 1640 (Thermo Fisher Scientific) media supplemented with 10% FBS (Sigma) as described previously (Giengkam et al 2015). Infected L929 cells were grown in the same way, but at 35 ^oC. Frozen stocks of bacteria were grown for 5 days, then the bacterial content was calculated using qPCR against the bacterial gene TSA47 (Giengkam et al., 2015). Bacteria were isolated onto fresh L929 cells in 75 cm^2 flasks at an Multiplicity of Infection of 10:1 and then grown for an additional 7 days. At this point bacteria were isolated from host cells and prepared for DNA extraction.

DNA extraction

The supernatant was removed from infected flasks and replaced with 6-8 ml pre-warmed media. Infected cells were harvested by mechanical scraping and then lysed using a bullet blender (BBX24B, Bullet Blender Blue, Nextadvance USA) operated at power 8 for 1 min. Host cell debris was removed by centrifugation at 300xg for 3 minutes, and the supernatant was filtered through a 2.0 µm filter unit. 10 µl of 1.4 µg/µl DNase (Deoxyribonuclease I from bovine pancreas, Merck, UK) was added per 1 ml of bacterial solution, then incubated at room temperature for 30 minutes. This procedure removed excess host cell DNA. The bacterial sample was then isolated by centrifugation at 14,000xg for 10 min at 4 ^oC, and washed two times with 0.3M sucrose (Merck, UK). After the washing steps were completed DNA was extracted using a QIAGEN Dneasy Blood & Tissue Kit (QIAGEN, UK) following the manufacturer’s instructions.

Purified DNA samples were analysed by gel electrophoresis using 0.8% agarose gel, in order to assess the DNA integrity. The yield of genomic DNA was quantified using a nanodrop (Nanodrop^{TM} 2000, Thermo Scientific, UK) and Qubit Fluorometric Quantitation (Qubit^{TM} 3.0 Fluorometer, Thermo Scientific, UK).

Sequencing

SMRTBell templates were prepared from purified *Orientia* genomic DNA according to PacBio's recommended protocols. Briefly, 20kb libraries were targeted; enrichment for large fragments was done using BluePippin (Sage Science) size selection method or successive Ampure (Beckman Coulter) clean-ups, depending on the original DNA size distribution and quantity, as recommended by PacBio. SMRTBell templates were sequenced on a Pacific Biosciences RSII Sequencer using P6 chemistry with a 240min run time. An average of 1.05 Gb of raw sequence was collected per strain (range 0.3-2.4 Gb), with an average N50 read length of 28.5 Kb (range 10.6-41.5 Kb). Genomes were assembled using the RS_HGAP_Assembly.3 protocol from the PacBio SMRTPortal (version 2.3.0), with initial polishing performed on trimmed initial assemblies using the same raw sequencing data with the RS_Resequencing.1 protocol. Each assembly was further polished using paired-end reads sequenced on an Illumina Miseq machine. Sequencing information and Illumina data availability for each sample can be found in Table S1; PacBio data is available under EBI accession PRJEB24834. For each assembly, the corresponding Illumina reads were aligned to the PacBio assembly using Stampy v1.0.23 (Lunter and Goodson, 2010). Pilon v1.16 (Walker et al., 2014) was then used to generate a final genome, and corrected 2 to 265 errors in the assemblies, with the majority of the errors being single base deletions at the end of A or T homopolymer runs. All genomes were rotated and reverse complemented as needed so that the predicted start codon for the dnaA gene formed the first nucleotide in the genome sequence. Sequencing and assembly statistics can be found in Table 2.

The Boryong, Ikeda, and non-*Orientia* Rickettsial genomes used in this study were obtained from NCBI (Table S2).

The finished assemblies were annotated using Prokka v1.11 (Seemann, 2014) , using a custom database created from the Boryong and Ikeda strains, which were previously annotated using the NCBI prokaryotic annotation pathway. The Boryong and Ikeda strains were re-annotated using Prokka for consistency with the other samples. Short gene names for all non-hypothetical gene products were checked manually (607 products). Where genes names were present for Boryong and/or Ikeda a consensus name based on these was selected. Where no short name was available, the long gene name was searched for in *E. coli* using the UniProt database, and where a single and unambiguous match was selected this was used. In cases of ambiguity the protein sequence from *Orientia* was used in a BLAST search against *E. coli, R. rickettsii* and *H. sapiens* and the short name of the closest match was selected. The key *Orientia* genes *TSA56, TSA47, TSA22, ScaA, ScaC, ScaD,* and *ScaE* were also manually annotated by taking known protein sequences from the UT76 strain and using BLAST to find the homologous genes in the other strains and give them the correct names. The single contig genomes were rotated to begin with the *DnaA* gene.

Repetitive regions of the genome were defined as regions of at least 1000bp in length which had a match with another 1000bp region with up to 100 differences (mismatches, insertions, and deletions) allowed. The repetitive regions were identified with Vmatch (Abouelhoda et al., 2004).

The core and accessory genome was identified using Roary (Page et al., 2015) with a threshold of 80% sequence identity required to consider two sequences part of the same gene group. Core genes were defined as genes present in every sample and as a single copy in every sample. The accessory genes identified using Roary were re-clustered using CD-Hit (Fu et al., 2012; Li and Godzik, 2006) with a cutoff of 80% identity across 95% of the length of the shortest protein to identify accessory genes which were truncated copies of other proteins. The correlation between gene order in each pair of samples was calculated by taking the order of the genes relative to the Karp strain and calculating the Spearman's rank coefficient between each pair. COG categories were assigned using RPS-BLAST to find matches in the NCBI Conserved Domain Database (Marchler-Bauer et al., 2002) and assigning a COG category to these using cdd2cog (Leimbach, 2016). Core repeat genes were identified using protein clusters generated by CD-Hit to find gene groups which were present at more than 1 copy. The clusters were identified using CD-Hit on the proteins predicted by Prokka with a cutoff of 80% identity across 90% of the length of the shortest protein. Pseudogenes were identified from CD-Hit protein clusters where at least one protein was a truncated version of the longest protein in the group. As pseudogenes which are truncated at the 5' end will not be annotated by Prokka, BLAST (Altschul et al., 1997) was using to screen for any additional pseudogenes in non-genic regions by searching for BLAST hits with protein identity >= 80% and an E-value $<10^{-15}$. This method found a further 26-37 pseudogenes per strain.

Further analysis used BioPython (Cock et al., 2009) and the GenomeDiagram package (Pritchard et al., 2006). Figure 1 was created with Circos (Krzywinski et al., 2009). Statistical tests were carried out in R (R Core Team, 2014) and the Python SciPy library (Jones et al.). Phylogenies were inferred using Maximum Likelihood methods in RaxML (Stamatakis, 2014) under the GAMMA model of rate heterogeneity and bootstrap values calculated using the rapid bootstrap method. The input sequences were aligned with Clustal Omega (Sievers et al., 2011) (for the 56kDa/46kDa/MLST trees) or using the MAFFT alignments produced by Roary (for the core gene tree). Phylogenetic trees were drawn using the ape (Paradis et al., 2004) and phytools (Revell, 2012) R packages, and Robinson-Foulds distances were calculated using the phangorn (Schliep, 2011) R package.

Results

Sequencing, Assembly, and Annotation

Eight genomes were assembled using the PacBio reads to perform initial genome assembly and Illumina sequencing reads to polish the genomes and reduce errors. Six of the eight genomes could be assembled into a single finished contig, while two genomes remain in multiple contigs. In addition, two previously assembled references genomes, *Orientia tsutsugamushi* str. Boryong and *Orientia tsutsugamushi* str. Ikeda, were incorporated into our analysis. The genome size ranges from 1.93Mb to 2.47Mb, and the GC content for all strains is consistent at 30-31%. We assessed the genomes to identify core genes shared between all genomes, and look for repetitive regions and repeat genes in each strain. Figure 1 plots the genetic elements of each complete genome.

The number of predicted genes in each strain ranges from 2086 (UT176) to 2709 (Gilliam) and is highly correlated with genome size (Spearman's correlation coefficient 0.94, $p < 2.2x10^{-16}$). A function could not be assigned, by similarity to reported sequences, to 325-547 genes (16 to 22 % of the identified coding regions) in each strain.

Core genome analysis

The set of 8 complete, single-contig genomes was used to identify core genes (present in all genomes) and accessory genes (present in a subset of genomes), using the criterion that all members of a group of putative orthologues should be at least 80% identical (similar) to all other members of the group. While the unfinished genomes do not appear to have lower numbers of predicted genes, which might indicate the assembly is incomplete, for this analysis the two strains which assembled as multiple contigs were excluded to avoid excluding core genes which are missing from the unfinished assemblies. A total of 657 gene groups were present in all 8 strains and therefore form a putative core genome, while 2812 gene groups were present in 2-7 of the 8 strains, and a further 4687 gene groups were found in a single strain. The 657 core genes make up 28-35% of the genome of each strain (Table S3). The number of core gene groups does not continue to decrease as more genomes are added to the analysis, suggesting that the core genome of *Orientia* can be defined with 8 representative genomes. In the initial analysis with Roary, the total number of gene groups continues to grow, suggesting an open pan-genome, but observation of the 7499 accessory gene groups showed that of the 6050 groups where a function can be assigned to one or more gene, there are only 122 distinct gene products, many of them conjugal transfer proteins, transposases, DNA helicases, and other functions shared by genes known to be part of the *Orientia tsutsugamushi* amplified genetic element identified in the Ikeda strain (Nakayama et al., 2008). Re-clustering these accessory genes but allowing genes which are only a match to part of a gene sequence to cluster together to include more truncated and fragmented copies of genes shows that the number of accessory gene groups continues to increase, but at a slower rate (Figure S1). The number of gene products remains constant at 122 no matter how many strains are included in the analysis. This suggests that the increase in non-core gene clusters is mainly due to further duplication and truncation of existing genes, rather than by the import of novel genes.

Genome Synteny

With the completed genomes produced by long read sequencing, the synteny of the genomes can be investigated. Previous work on the Boryong and Ikeda genomes showed extensive genome shuffling between the two strains. Analysis of the order and grouping of the core genes which are conserved in each genome shows that the genome has undergone massive rearrangement, with the core genes found in core gene 'islands' with repeat regions interspersed between these islands. The 657 core genes are present in 145-157 distinct islands, of which only 51 are conserved (defined as the same genes present in the same order) in all genomes. Figure 3 shows the position and ordering of these conserved core gene islands which are maintained in all samples relative to the position and ordering in the Karp strain. The correlation between gene order in each pair of samples is shown in Figure S2. A value close to 0 shows low correlation in gene order, while values closer to 1 show higher correlation in gene order. As there are differences in the correlation of gene order between strains, this suggests that the process of genome rearrangement is happening in multiple steps and not as a single event.

The identities of genes present on conserved islands is shown in table S5. Conserved islands range from 1-13 genes in size, with larger islands often containing genes linked by plausible biological functions. For example, groups 3 and 6 include a number of cell division and peptidoglycan biosynthesis genes (including *mraY, murF, murE, pbp, ftsL, dnaJ* and *dnaK* in group 3 and *murC, murB, ddl* and *ftsQ* in group 6) and groups 31 and 32 include a number of 30S and 50S ribosomal proteins. Analysis of the number of conserved islands shared between samples shows that the number of conserved islands continues to decrease as more genomes are included (Figure S3), and suggests that gene order and clustering is not always constrained in *Orientia tsutsugamushi*. There is no difference seen in the size of the islands between conserved and non-conserved islands (Figure S4) (two-sample Kolmogorov-Smirnov test D=0.085, p-value=0.86), the nucleotide diversity between genes in the two categories of islands (Figure S5) (two-sample Kolmogorov-Smirnov test D=0.052, p-value=0.86), or the Clusters of Orthologous Groups (COG) categories assigned to genes in the two island categories (Chi-squared test χ^2= 15.03, p=0.82).

Repeats and pseudogenes

The genomes of *Orientia tsutsugamushi* are known to be highly repetitive, including a highly amplified genetic element known as the *Orientia tsutsugamushi* amplified genetic element (Otage), as well as other transposable elements.

Our results emphasise the large number of repeated genes and regions, including many genes related to the Type IV secretion system. The total proportion of the genome which is repetitive (see Methods for our definition of repetitive) differs markedly from 33% in UT176 to 51% in Gilliam (Table S3). In contrast, the extremely compact (and therefore non-repetitive) *Rickettsia typhi* genome is 0% repetitive by our measure and even, intriguingly, the *Rickettsia* endosymbiont of *Ixodes scapularis*, known to encode multiple copies of the same repetitive element found in *Orientia* (Gillespie et al., 2012), is 20% repetitive in our analysis, despite our methods giving more conservative figures than previously determined for the Ikeda strain (Nakayama et al., 2008).

We identified 530 groups of repeat genes containing 12043 genes present in multiple copies in at least one strain, which we term "core repeats". Of the 530 groups, 427 represent genes found in multiple strains, which 103 are found only in a single strain. Despite clustering in 530 groups, the genes have only 66 different functional products, as is expected from the earlier results looking at all the non-core genes. The repeat genes are mainly transposase and conjugal transfer genes, similar to those previously reported in the Otage (Table S6), and cluster into genetic elements which are interspersed between the core genes. Many of these genes are present in high copy number, with all strains carrying over 200 transposases and 300 conjugal transfer genes and gene fragments. These core repeat elements occupy 35-47% of the *Orientia tsutsugamushi* genome and represent 57-67% of the genes in these genomes (Table S4).

Orientia tsutsugamushi genes are known to exhibit high levels of pseudogenisation and gene decay. We searched for pseudogenes in each genome, and identified up to 484 pseudogenes per strain (Table S7). This is lower than previously reported in Ikeda, but due to methodological differences the figures cannot be directly compared. We also assessed whether the pseudogene had been caused by truncation at the 5' or 3' end of the sequencing, or by frameshift.

Phylogenetics

A phylogenetic tree was constructed using the core genes from each strain. This can be compared to trees built using the 56kDa (Figure 4) and 47kDa (Figure S6) genes, which are often used for phylogenetic analysis of *Orientia tsutsugamushi,* or to trees built using the

MLST genes (Figure S7). *Orientia* strains are commonly based on their similarity to reference strains, either from phylogenetics or serology. Compared to the 56kDa tree, the core gene tree suggests the Kato and Ikeda strains are more closely related to the Karp, UT176, and UT76 strains than the TA686 and Gilliam strains (Figure 4). Robinson-Foulds distances between trees are shown in Table S8; for this small number of strains, the distance is lowest between the 47kDa tree and the core genome tree.

Discussion

We present the first large-scale study of *Orientia tsutsugamushi*, a bacterium which is important both for the study of human disease and for its unique insights into genome evolution.

Previous studies of *Orientia tsutsugamushi* genomes have used BAC cloning and Sanger sequencing to produce complete genomes (Cho et al., 2007; Nakayama et al., 2008), or have used next-generation sequencing strategies which have produced only incomplete and fragmented genomes (Liao et al., 2017). We demonstrate that a combination of PacBio and Illumina sequencing is sufficient to produce a single-contig genome, allowing us to study the gene content and synteny of this organism. For the two genomes which could not be assembled into single contigs in our study (FPW1038 and TA763), we found that the sequencing produced fewer reads at the high end of the length distribution. This suggests that given the highly repetitive nature of the *Orientia tsutsugamushi* genome, the DNA preparation and sequencing methods must be carefully chosen to produce very long reads in order to produce complete assemblies. We used Illumina sequencing to correct errors in our genomes, which was vital to reduce the number of homopolymer errors, which could otherwise suggest frameshift errors and affect gene annotation. While the fewest errors we corrected in a strain was two, this is likely an underestimate as errors in repetitive regions where Illumina reads cannot map are impossible to correct. While our analysis shows small differences when quantifying the extent of the repeat regions and repeat gene families in *Orientia* compared to previous work, a direct comparison is difficult due to differences in methodology between analyses.

Owing to the difficulties of producing complete genomes, most previous work has relied on single gene or MLST studies to investigate the genetic diversity of *Orientia tsutsugamushi.* We demonstrate that phylogenies generated from limited data are substantially different from those produced from the whole core genome. The common practice of grouping *Orientia* strains into 'Karp-like' or 'Gilliam-like' groups based on the genotype of the 56kDa antigen may not give an accurate view of the relatedness of these strains, especially when recombination is taken into account, although this may still be important when considering immune response.

Previous work has demonstrated limited synteny between the two reference strains of *Orientia tsutsugamushi*, but we extend this to demonstrate that there is minimal synteny between any known *Orientia tsutsugamushi* genome. The pattern of core gene islands separated by transposable elements and repeats suggests a repeat-mediated system of chromosome rearrangement. It is unclear whether this is a gradual process of genome rearrangement, or whether the genome is being broken apart and rearranged rapidly, similar to chromothripsis or the chromosome repair of *Deinoccocus radiodurans* after exposure to ionizing radiation. In *Deinococcus*, it is thought that RecFOR pathway is particularly important for DNA repair, and it has no homologues to RecB or RecC (Cox et al., 2010). Similarly, in *Orientia*, the core genome does not contain RecB or RecC, but does contain the RecFOR pathway genes, indicating this alternative DNA repair pathway may also be important. Longitudinal studies of *Orientia tsutsugamushi* genomes during passage or infection may be needed to determine the speed and processes of genome rearrangement in *Orientia*.

We report a core genome of only 657 genes, compared to the 519 previously reported as the core genome shared between *Orientia* and five other sequenced *Rickettsia*, despite using a relatively low sequence identity threshold to determine gene clusters. Differences in methodology may lead to the reporting of different core gene sets, but more interesting is the pattern of core genome islands separated by amplified repeat regions, and the lack of conservation in the ordering and clustering of the core genes.

All of the *Orientia* genomes show high repetitiveness, which we measured as both non-unique regions of the genome, and genes which are present in multiple copies (some of which may be truncated). The genomes of intracellular bacteria tend towards genome reduction and gene loss (Darby et al., 2007; Merhej and Raoult, 2011), but maintain degraded genes and accumulate non-coding DNA. The transition to intracellularity has been hypothesized to lead to the relaxation of selective pressure on the genome (Moran, 1996), with an increased rate of sequence evolution. The expansion of the Otage (and other mobile elements) throughout the *Orientia* lineage appears to be another consequence of relaxed selection on *Orientia* in its intracellular niche, again leading to accelerated sequence evolution of the genome through rearrangement and gene loss. This is supported by the finding that the diversity of gene repertoire between strains of *Orientia tsutsugamushi* is largely due to the duplication and truncation of existing genes, and we find no evidence for the acquisition of new genetic material via horizontal transfer . The amplication of a transposable element has been seen in Rickettsial (Gillespie et al., 2012) and non-Rickettsial (Wiens et al., 2008) species, but it is not known whether this is associated with rearrangement of the genome in other species.

In conclusion, we report the generation of six complete and a further two draft genomes from a diverse set of strains of the important but neglected human pathogen *Orientia tsutsugamushi.* This set includes the major reference strains Karp, Kato and Gilliam, and will serve as a valuable resource for scientists and clinicians studying this pathogen, in particular supporting future work on *Orientia* genomics, vaccine development, and cell biology. The new genomes reported here confirm the status of *Orientia* as one of the most fragmented and highly repeated bacterial genomes known, and exciting questions remain regarding the mechanisms and timeframes driving the evolution of these extraordinary genomes.

Conflict of Interest

P.D. is Founder, Director, and Executive Officer of Genomics plc and a Partner of Peptide Groove LLP.

Data Availability

Sequence data and assemblies generated in this study have been uploaded to the EBI under project PRJEB24834.

Author Contributions

Acknowledgements and Funding

J.S. was funded by a Royal Society Dorothy Hodgkin Research Fellowship. P.D. is supported by a Wellcome Trust Core Award (090532/Z/09/Z).

References

Abouelhoda, M.I., Kurtz, S., and Ohlebusch, E. (2004). Replacing suffix trees with enhanced suffix arrays. J. Discret. Algorithms *2*, 53–86.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped {BLAST} and {PSI-BLAST:} a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Arai, S., Tabara, K., Yamamoto, N., Fujita, H., Itagaki, A., Kon, M., Satoh, H., Araki, K., Tanaka-Taya, K., Takada, N., et al. (2013). Molecular phylogenetic analysis of Orientia tsutsugamushi based on the groES and groEL genes. Vector Borne Zoonotic Dis. *13*, 825–829.

Blacksell, S.D., Luksameetanasan, R., Kalambaheti, T., Aukkanit, N., Paris, D.H., McGready, R., Nosten, F., Peacock, S.J., and Day, N.P.J. (2008). Genetic typing of the 56-kDa type-specific antigen gene of contemporary *Orientia tsutsugamushi* isolates causing human scrub typhus at two sites in north-eastern and western Thailand. FEMS Immunol. Med. Microbiol. *52*, 335–342.

Bonell, A., Lubell, Y., Newton, P.N., Crump, J.A., and Paris, D.H. (2017). Estimating the burden of scrub typhus: A systematic review. PLoS Negl. Trop. Dis. *11*, e0005838.

Cho, N.-H., Kim, H.-R., Lee, J.-H., Kim, S.-Y., Kim, J., Cha, S., Kim, S.-Y., Darby, A.C., Fuxelius, H.-H., Yin, J., et al. (2007). The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. Proc. Natl. Acad. Sci. U. S. A. *104*, 7981–7986.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Coleman, R.E., Monkanna, T., Linthicum, K.J., Strickman, D.A., Frances, S.P., Tanskul, P., Kollars, T.M., Inlao, I., Watcharapichat, P., Khlaimanee, N., et al. (2003). Occurrence of Orientia tsutsugamushi in small mammals from Thailand. Am. J. Trop. Med. Hyg. *69*, 519–524.

Cox, M.M., Keck, J.L., and Battista, J.R. (2010). Rising from the Ashes: DNA Repair in Deinococcus radiodurans. PLoS Genet. *6*, e1000815.

Darby, A.C., Cho, N.-H., Fuxelius, H.-H., Westberg, J., and Andersson, S.G.E. (2007). Intracellular pathogens go extreme: genome evolution in the Rickettsiales. Trends Genet. *23*, 511–520.

Dittrich, S., Rattanavong, S., Lee, S.J., Panyanivong, P., Craig, S.B., Tulsiani, S.M., Blacksell, S.D., Dance, D.A.B., Dubot-Pérès, A., Sengduangphachanh, A., et al. (2015). Orientia, rickettsia, and leptospira pathogens as causes of CNS infections in Laos: a prospective study. Lancet. Glob. Heal. *3*, e104-12.

Duong, V., Blassdell, K., May, T.T.X., Sreyrath, L., Gavotte, L., Morand, S., Frutos, R., and Buchy, P. (2013). Diversity of Orientia tsutsugamushi clinical isolates in Cambodia reveals active selection and recombination process. Infect. Genet. Evol. *15*, 25–34.

Enatsu, T., Urakami, H., and Tamura, A. (1999). Phylogenetic analysis of *Orientia tsutsugamushi* strains based on the sequence homologies of 56-kDa type-specific antigen genes. FEMS Microbiol. Lett. *180*, 163–169.

Frutos, R., Viari, A., Ferraz, C., Morgat, A., Eychenié, S., Kandassamy, Y., Chantal, I., Bensaid, A., Coissac, E., Vachiery, N., et al. (2006). Comparative genomic analysis of three strains of Ehrlichia ruminantium reveals an active process of genome size plasticity. J. Bacteriol. *188*, 2533–2542.

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

Giengkam, S., Blakes, A., Utsahajit, P., Chaemchuen, S., Atwal, S., Blacksell, S.D., Paris, D.H., Day, N.P.J., and Salje, J. (2015). Improved Quantification, Propagation, Purification and Storage of the Obligate Intracellular Human Pathogen Orientia tsutsugamushi. PLoS Negl. Trop. Dis. *9*, e0004009.

Gillespie, J.J., Joardar, V., Williams, K.P., Driscoll, T., Hostetler, J.B., Nordberg, E., Shukla, M., Walenz, B., Hill, C.A., Nene, V.M., et al. (2012). A Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. J. Bacteriol. *194*, 376–394.

Izzard, L., Fuller, A., Blacksell, S.D., Paris, D.H., Richards, A.L., Aukkanit, N., Nguyen, C., Jiang, J., Fenwick, S., Day, N.P.J., et al. (2010). Isolation of a Novel *Orientia* Species (*O. chuto* sp. nov.) from a Patient Infected in Dubai. J. Clin. Microbiol. *48*, 4404–4409.

Jiang, J., Paris, D.H., Blacksell, S.D., Aukkanit, N., Newton, P.N., Phetsouvanh, R., Izzard, L., Stenos, J., Graves, S.R., Day, N.P.J., et al. (2013). Diversity of the 47-kD HtrA nucleic acid and translated amino acid sequences from 17 recent human isolates of Orientia. Vector Borne Zoonotic Dis. *13*, 367–375.

Jones, E., Oliphant, T., Peterson, P., and others {SciPy}: Open source scientific tools for {Python}.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Leimbach, A. (2016). bac-genomics-scripts: Bovine E. coli mastitis comparative genomics edition.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Liao, H.-M., Chao, C.-C., Lei, H., Li, B., Tsai, S., Hung, G.-C., Ching, W.-M., and Lo, S.-C. (2017). Intraspecies comparative genomics of three strains of Orientia tsutsugamushi with different antibiotic sensitivity. Genomics Data *12*, 84–88.

Lu, H.-Y., Tsai, K.-H., Yu, S.-K., Cheng, C.-H., Yang, J.-S., Su, C.-L., Hu, H.-C., Wang, H.-C., Huang, J.-H., and Shu, P.-Y. (2010). Phylogenetic analysis of 56-kDa type-specific antigen gene of Orientia tsutsugamushi isolates in Taiwan. Am. J. Trop. Med. Hyg. *83*, 658–663.

Luce-Fedrow, A., Lehman, M., Kelly, D., Mullins, K., Maina, A., Stewart, R., Ge, H., John, H., Jiang, J., and Richards, A. (2018). A Review of Scrub Typhus (Orientia tsutsugamushi and Related Organisms): Then, Now, and Tomorrow. Trop. Med. Infect. Dis. *3*, 8.

Lunter, G., and Goodson, M. (2010). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. *30*, 281–283.

McGready, R., Blacksell, S.D., Luksameetanasan, R., Wuthiekanun, V., Jedsadapanpong, W., Day, N.P.J., and Nosten, F. (2010). First report of an Orientia tsutsugamushi type TA716-related scrub typhus infection in Thailand. Vector Borne Zoonotic Dis. *10*, 191–193.

Merhej, V., and Raoult, D. (2011). Rickettsial evolution in the light of comparative genomics. Biol. Rev. Camb. Philos. Soc. *86*, 379–405.

Moran, N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A. *93*, 2873–2878.

Nakayama, K., Yamashita, A., Kurokawa, K., Morimoto, T., Ogawa, M., Fukuhara, M., Urakami, H., Ohnishi, M., Uchiyama, I., Ogura, Y., et al. (2008). The Whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. DNA Res. *15*, 185–199.

Nakayama, K., Kurokawa, K., Fukuhara, M., Urakami, H., Yamamoto, S., Yamazaki, K., Ogura, Y., Ooka, T., and Hayashi, T. (2010). Genome comparison and phylogenetic analysis of Orientia tsutsugamushi strains. DNA Res. *17*, 281–291.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693.

Paradis, E., Claude, J., and Strimmer, K. (2004). A{PE}: analyses of phylogenetics and evolution in {R} language. Bioinformatics *20*, 289–290.

Paris, D.H., Aukkanit, N., Jenjaroen, K., Blacksell, S.D., and Day, N.P.J. (2009). A highly sensitive quantitative real-time PCR assay based on the groEL gene of contemporary Thai strains of Orientia tsutsugamushi. Clin. Microbiol. Infect. *15*, 488–495.

Paris, D.H., Phetsouvanh, R., Tanganuchitcharnchai, A., Jones, M., Jenjaroen, K., Vongsouvath, M., Ferguson, D.P.J., Blacksell, S.D., Newton, P.N., Day, N.P.J., et al. (2012). Orientia tsutsugamushi in human scrub typhus eschars shows tropism for dendritic cells and monocytes rather than endothelium. PLoS Negl. Trop. Dis. *6*, e1466.

Phetsouvanh, R., Sonthayanon, P., Pukrittayakamee, S., Paris, D.H., Newton, P.N., Feil, E.J., Day, N.P.J., Kurup, A., Issac, A., Loh, J., et al. (2015). The Diversity and Geographical Structure of Orientia tsutsugamushi Strains from Scrub Typhus Patients in Laos. PLoS Negl. Trop. Dis. *9*, e0004024.

Pritchard, L., White, J.A., Birch, P.R.J., and Toth, I.K. (2006). GenomeDiagram: a python package for the visualization of large-scale genomic data. Bioinformatics *22*, 616–617.

R Core Team (2014). R: A Language and Environment for Statistical Computing.

Revell, L.J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223.

Rights, F.L., and Smadel, J.E. (1948). Studies on scrub typhus; tsutsugamushi disease; heterogenicity of strains of R. tsutsugamushi as demonstrated by cross-vaccination studies. J. Exp. Med. *87*, 339–351.

Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592–593.

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

2069.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*.

Sonthayanon, P., Peacock, S.J., Chierakul, W., Wuthiekanun, V., Blacksell, S.D., Holden, M.T.G., Bentley, S.D., Feil, E.J., and Day, N.P.J. (2010). High rates of homologous recombination in the mite endosymbiont and opportunistic human pathogen Orientia tsutsugamushi. PLoS Negl. Trop. Dis. *4*, e752.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One *9*, e112963.

Weitzel, T., Dittrich, S., López, J., Phuklia, W., Martinez-Valdebenito, C., Velásquez, K., Blacksell, S.D., Paris, D.H., and Abarca, K. (2016). Endemic Scrub Typhus in South America. N. Engl. J. Med. *375*, 954–961.

Wiens, G.D., Rockey, D.D., Wu, Z., Chang, J., Levy, R., Crane, S., Chen, D.S., Capri, G.R., Burnett, J.R., Sudheesh, P.S., et al. (2008). Genome sequence of the fish pathogen Renibacterium salmoninarum suggests reductive evolution away from an environmental Arthrobacter ancestor. J. Bacteriol. *190*, 6970–6982.

Wongprompitak, P., Duong, V., Anukool, W., Sreyrath, L., Mai, T.T.X., Gavotte, L., Moulia, C., Cornillot, E., Ekpo, P., Suputtamongkol, Y., et al. (2015). Orientia tsutsugamushi, agent of scrub typhus, displays a single metapopulation with maintenance of ancestral haplotypes throughout continental South East Asia. Infect. Genet. Evol. *31*, 1–8.

Wu, M., Sun, L. V, Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J.C., McGraw, E.A., Martin, W., Esser, C., Ahmadinejad, N., et al. (2004). Phylogenomics of the Reproductive Parasite Wolbachia pipientis wMel: A Streamlined Genome Overrun by Mobile Genetic Elements. PLoS Biol. *2*, e69.

Figures and Tables

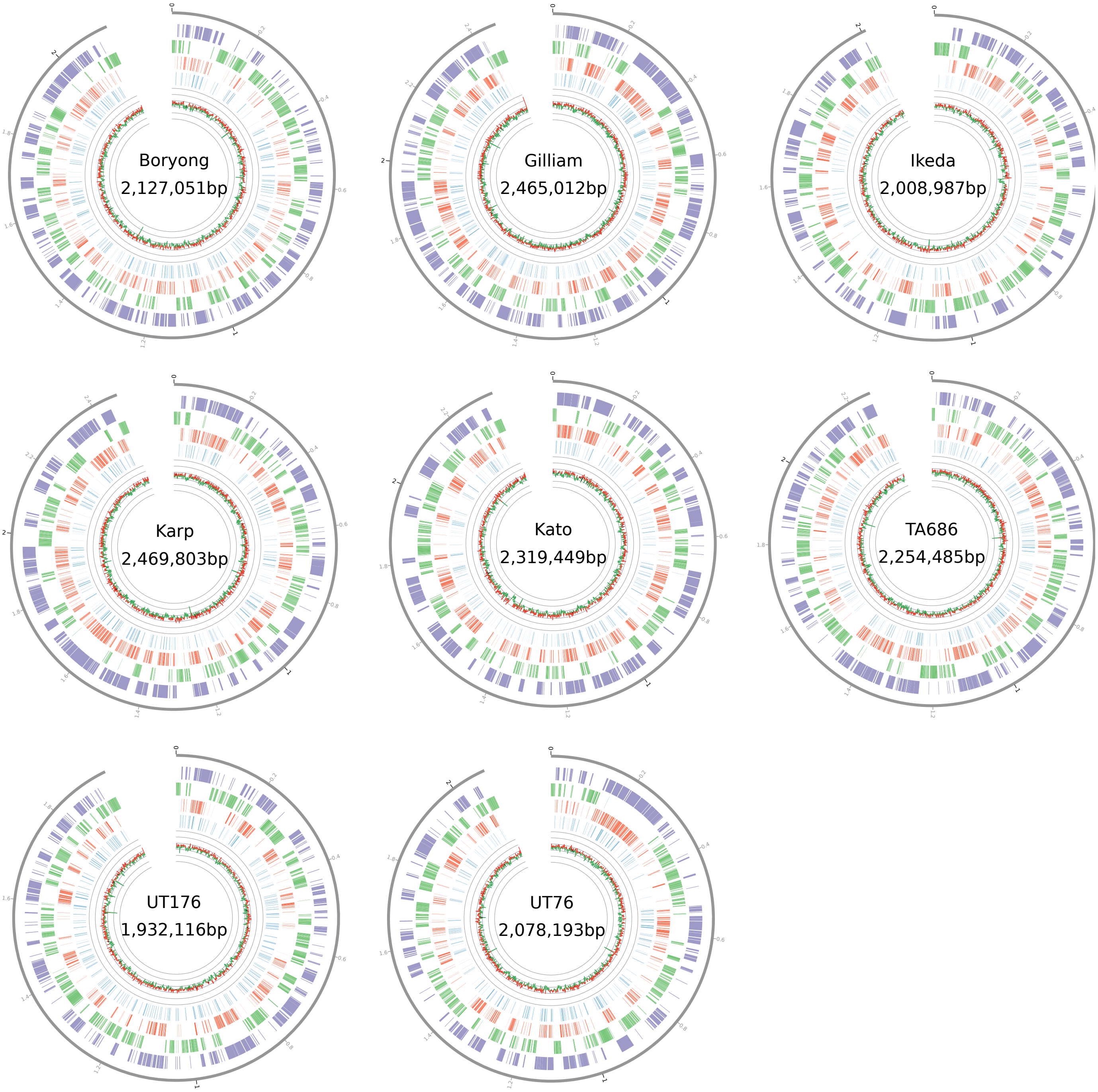
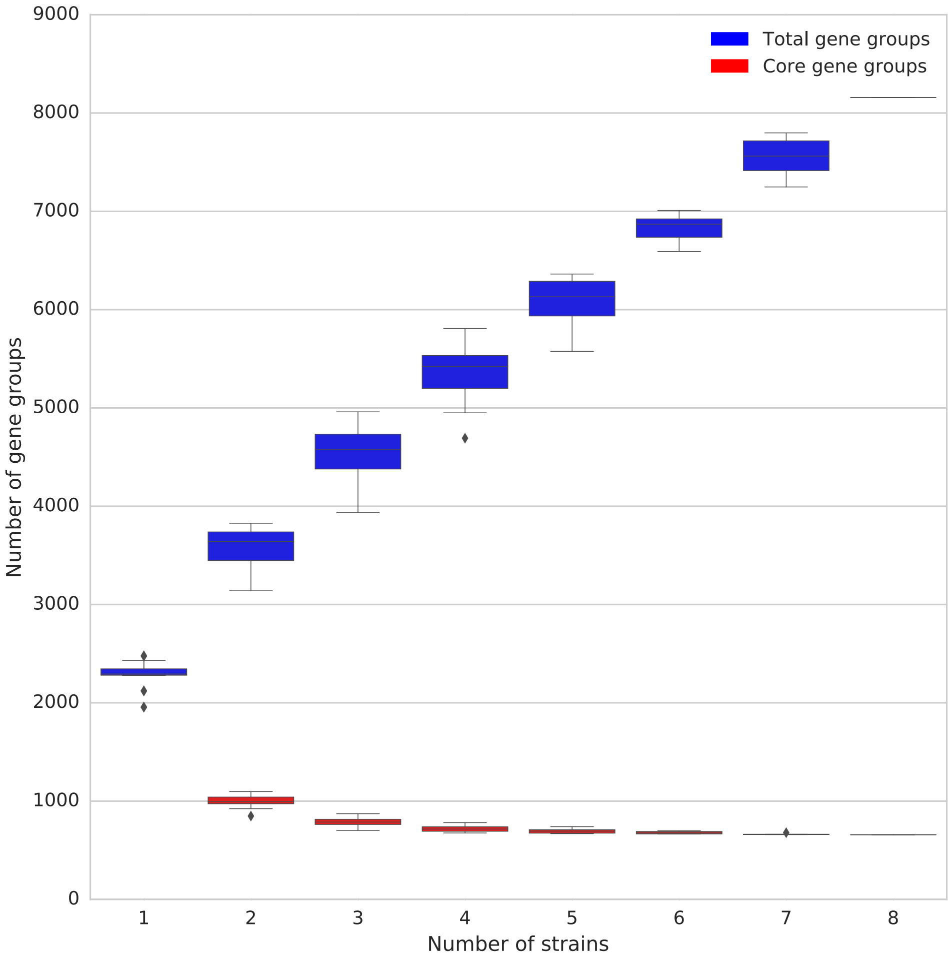


Figure 1. Ring diagrams for all single-contig strains. From outermost feature in each genome, moving inwards: repetitive regions are shown in purple, core genes in green, repeat genes in red and pseudogenes in blue. The track shows the GC percentage in windows of 1000bp. Values above the median GC are in green, and values below the median GC are in red.



630

631 *Figure 2: The number of core gene groups and the total number of gene groups (including*
632 *the core gene groups) as more strains are added to the analysis. Boxplots represent all*
633 *possible combinations of the number of strains given on the x-axis.*

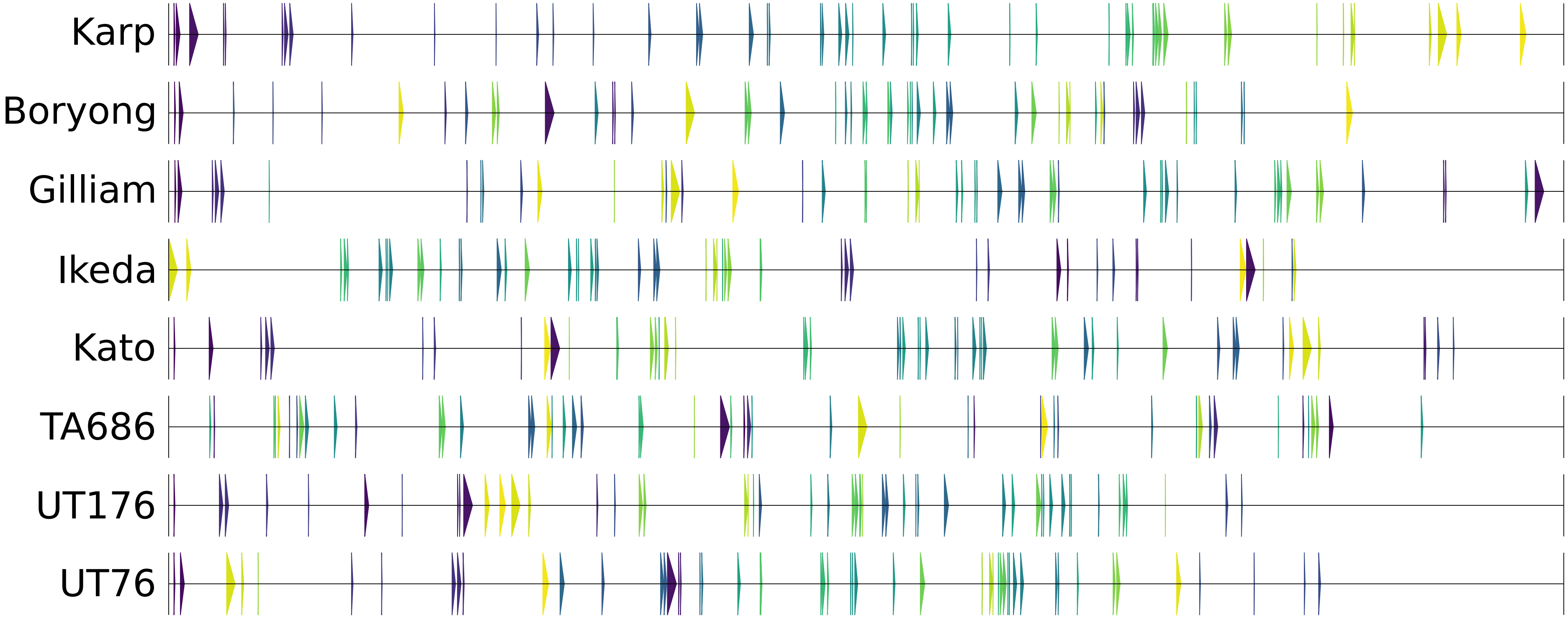


Figure 3. Each arrow represents the location of a core gene island containing one or more core genes which are conserved in the same order within an island across all strains. The arrows are coloured relative to their order in the Karp genome.

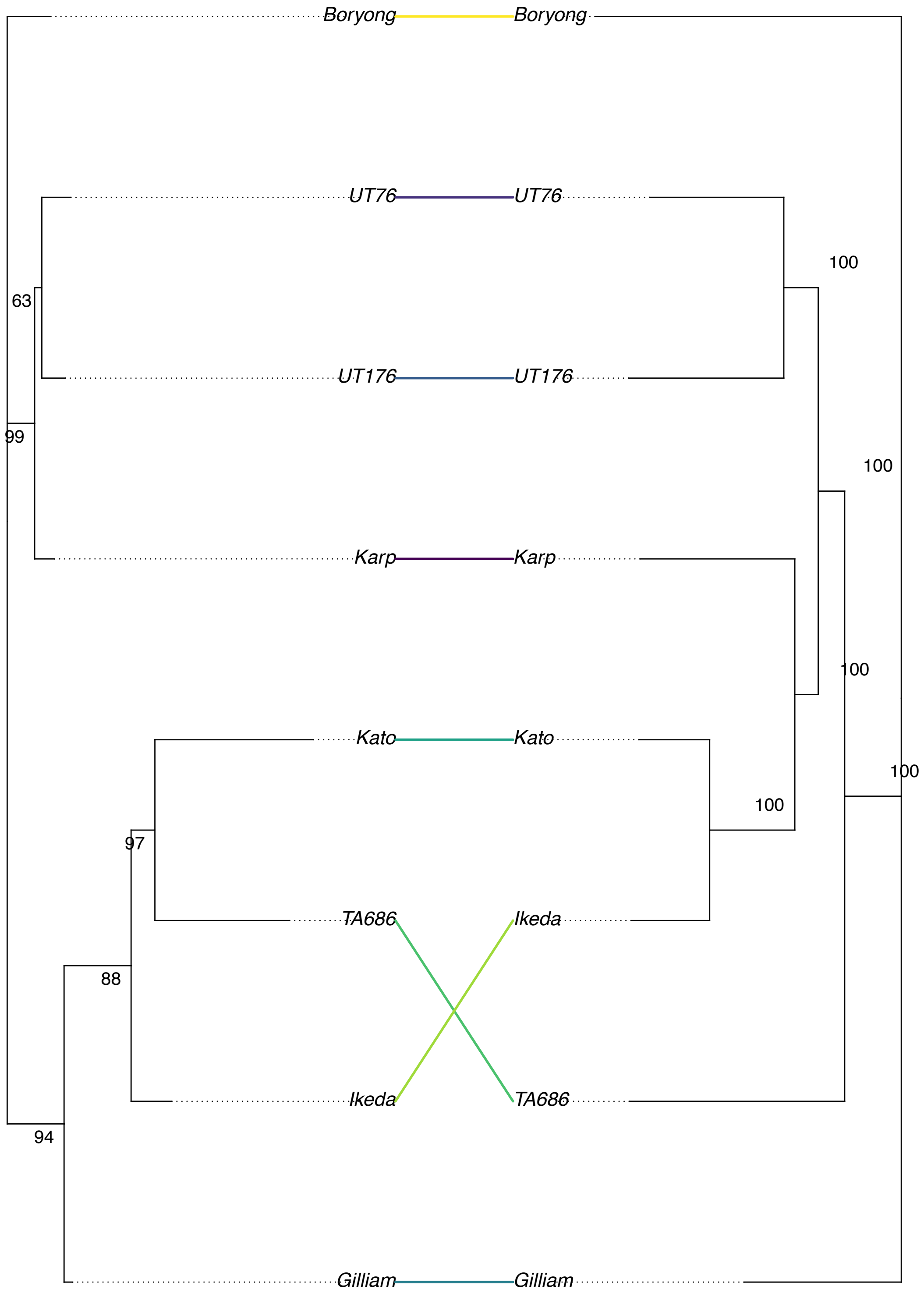


Figure 4. Phylogenetic trees generated from the 56kDa antigen sequence (left) and the sequence of the 657 core genes (right). The tree was inferred using the maximum likelihood method implemented in RaxML, and bootstrap values were calculated with the RaxML rapid bootstrap method.

Strain	**Original Source**	**Source in this study**	**Reference**
Karp	New Guinea, human patient, 1943	Naval Medical Research Centre (NMRC)	(Enatsu et al., 1999)
Kato	Niigata, Japan, human patient, 1955	NMRC	(Enatsu et al., 1999)
Gilliam	Indian-Burmese border, human patient, 1943	NMRC	(Rights and Smadel, 1948)
TA686	Thailand, animal (*Tupaia glis),* 1963	NMRC	(Enatsu et al., 1999)
TA763	Thailand, animal *Rattus rajah),* 1963	NMRC	(Enatsu et al., 1999)
FPW1038	Thailand-Burmese border, human patient (pregnant), 2010	Mahidol-Oxford Research Centre (MORU)	(McGready et al., 2010)
UT76	Udon Thani, Thailand, human patient, 2003	MORU	(Blacksell et al., 2008)
UT176	Udon Thani, Thailand, human patient, 2004	MORU	(Paris et al., 2009)

642 Table 1. Bacterial strains used in this study.

643

Strain	Genome length (bp)	Contigs	GC percentage	Errors corrected by Illumina sequencing
Boryong*	2,127,051	1	31	-
Ikeda*	2,008,987	1	31	-
FPW1038	2,035,338	25	31	265
Gilliam	2,465,012	1	31	7
Karp	2,469,803	1	31	48
Kato	2,319,449	1	31	5
TA686	2,254,485	1	31	28
TA763	2,089,396	8	31	88
UT76	2,078,193	1	30	2
UT176	1,932,116	1	30	13

Table 2. Assembly statistics for the 10 assemblies used in this analysis. Genomes marked with * are previously-assembled reference strains.

Strain	Genes	Annotated as hypothetical
Boryong	2443	547
Ikeda	2186	417
FPW1038	2198	369
Gilliam	2709	463
Karp	2578	470
Kato	2406	465
TA686	2546	474
TA763	2212	396
UT76	2247	420
UT176	2086	325

Table 3. Number of genes predicted in each strain after annotation with Prokka, and the number of genes which were annotated as hypothetical. The Boryong and Ikeda strains were reannotated with Prokka for consistency between strains.

Supplementary Figures and Tables

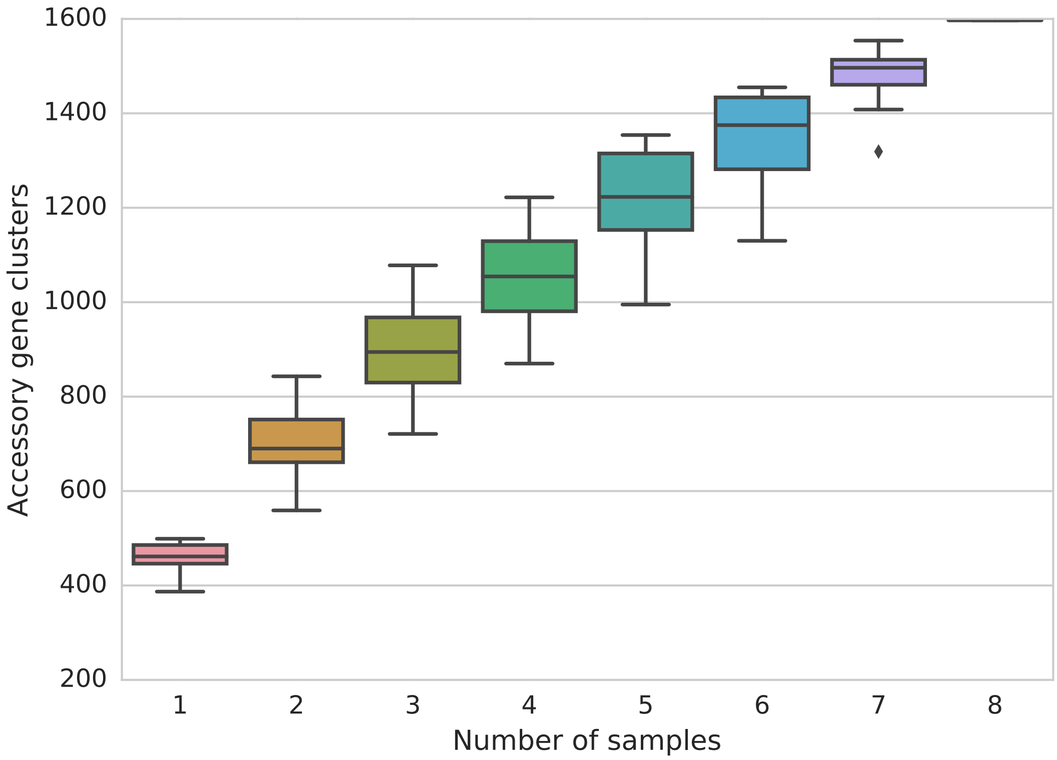


Figure S1. Boxplot of accessory genes clustered with a lenient length threshold to show how the number of clusters increases with number of samples included in the analysis.

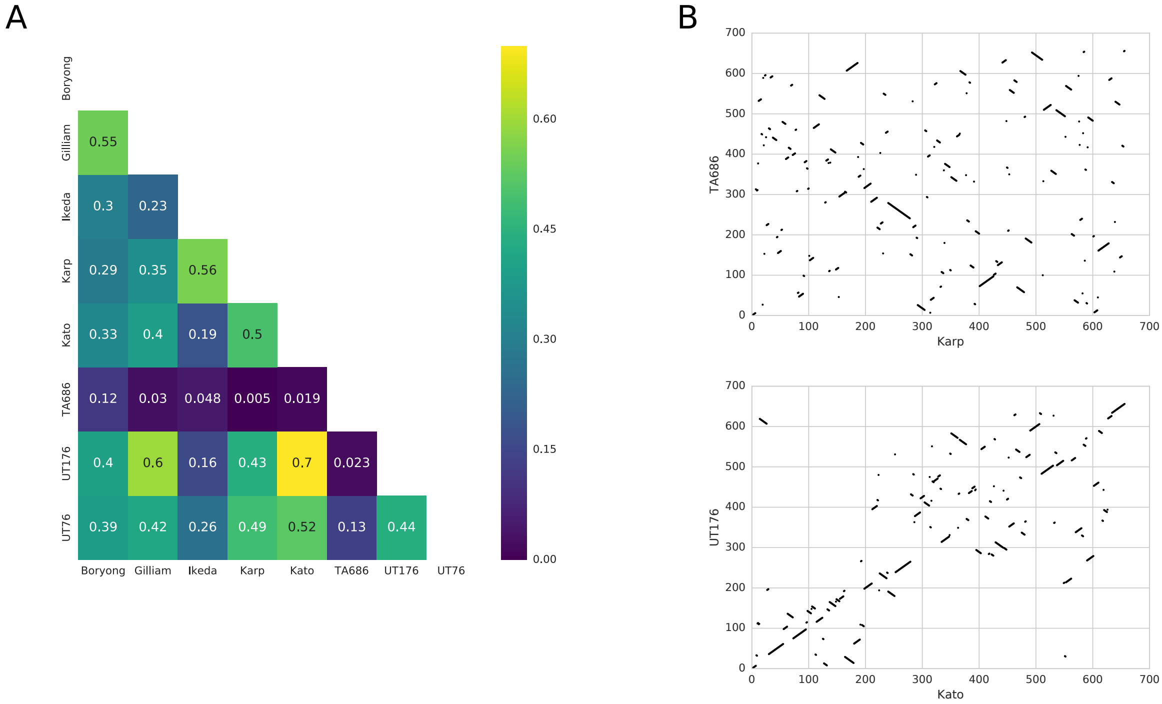


Figure S2. A - Heatmap showing the correlation in gene order between each pair of samples. B – dotplots showing the gene ordering between the pair with the highest correlation (Kato and UT176) and the lowest correlation (Karp and TA686).

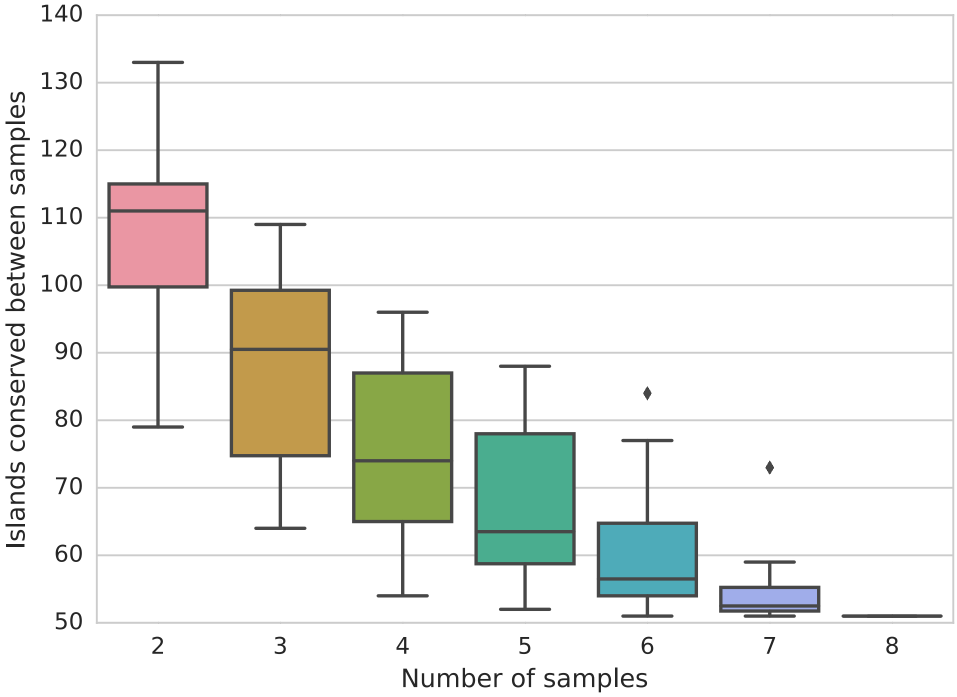
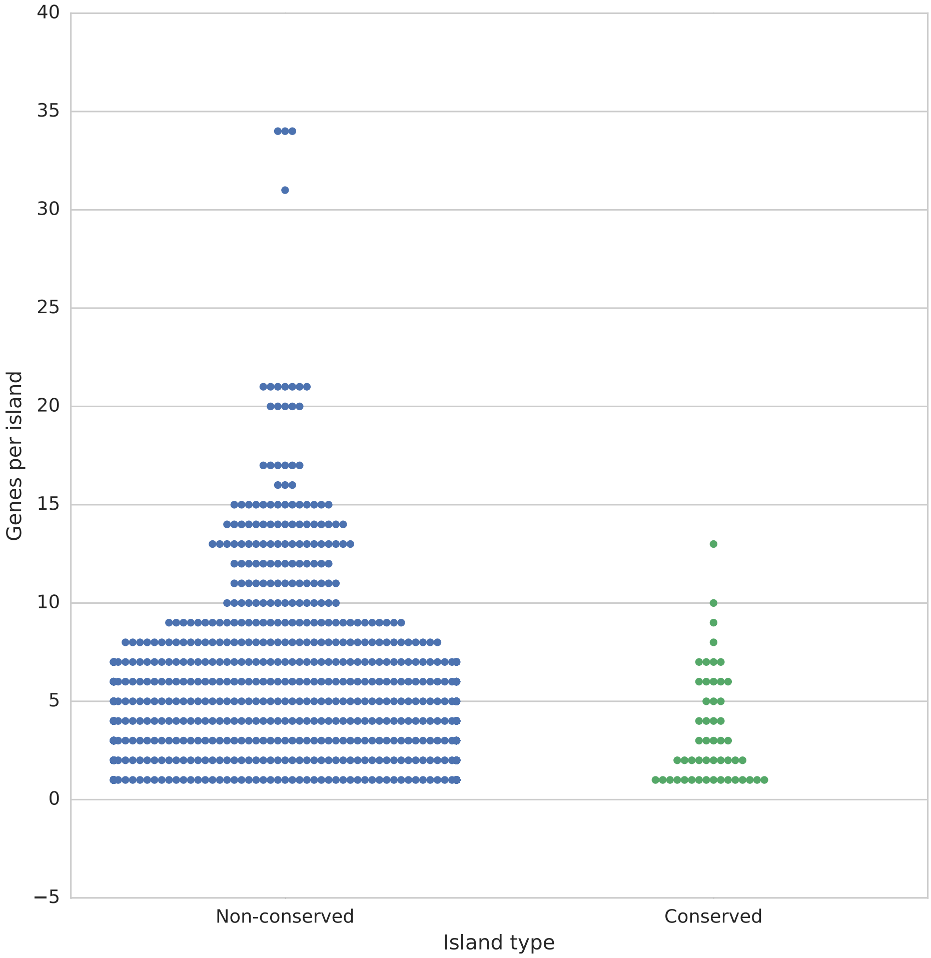


Figure S3. Boxplot showing the number of islands conserved between samples across all different combinations of samples.



669

670 *Figure S4. The number of genes per island in conserved versus non-conserved islands.*

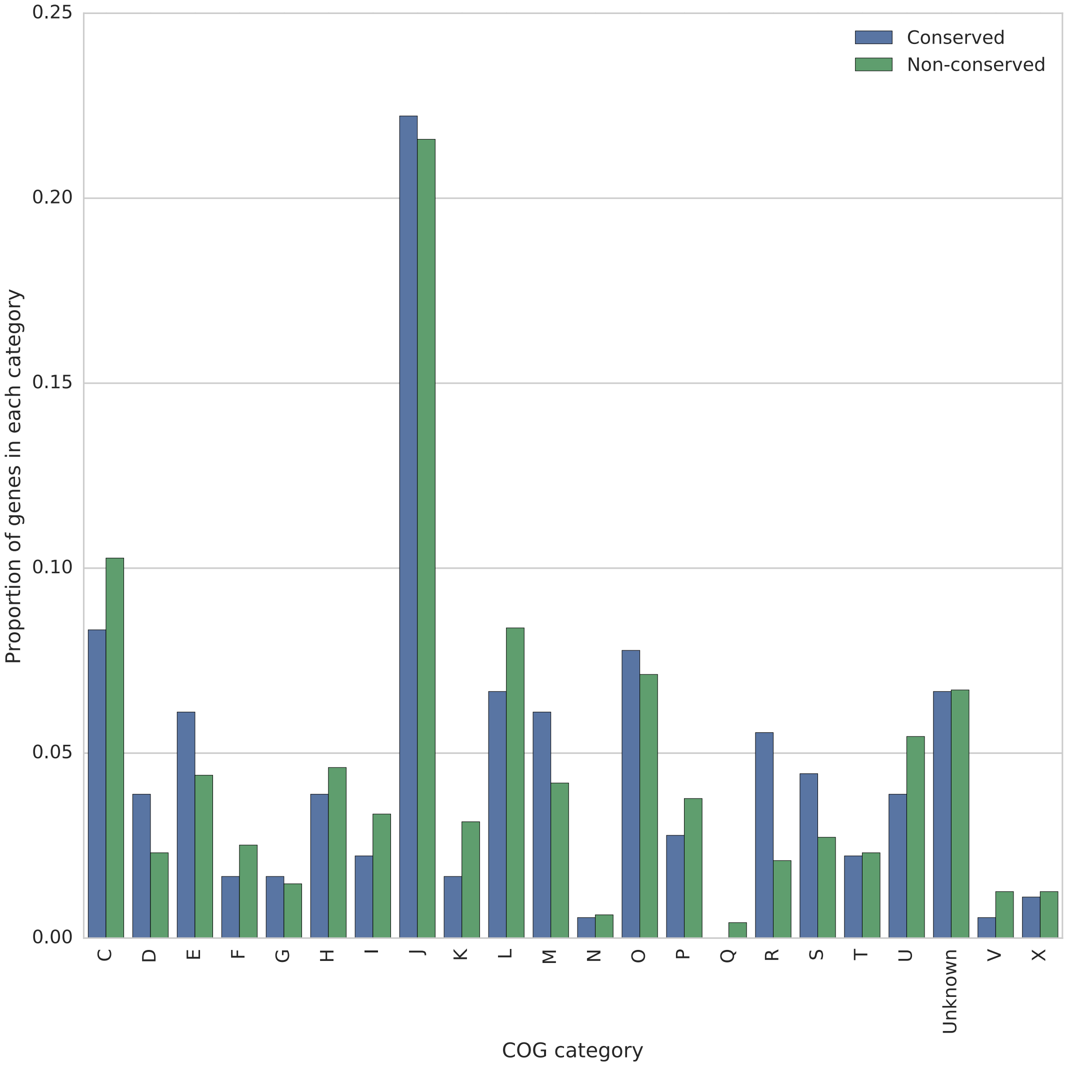


Figure S5. The proportion of core genes which are in conserved and non-conserved islands in each COG category.

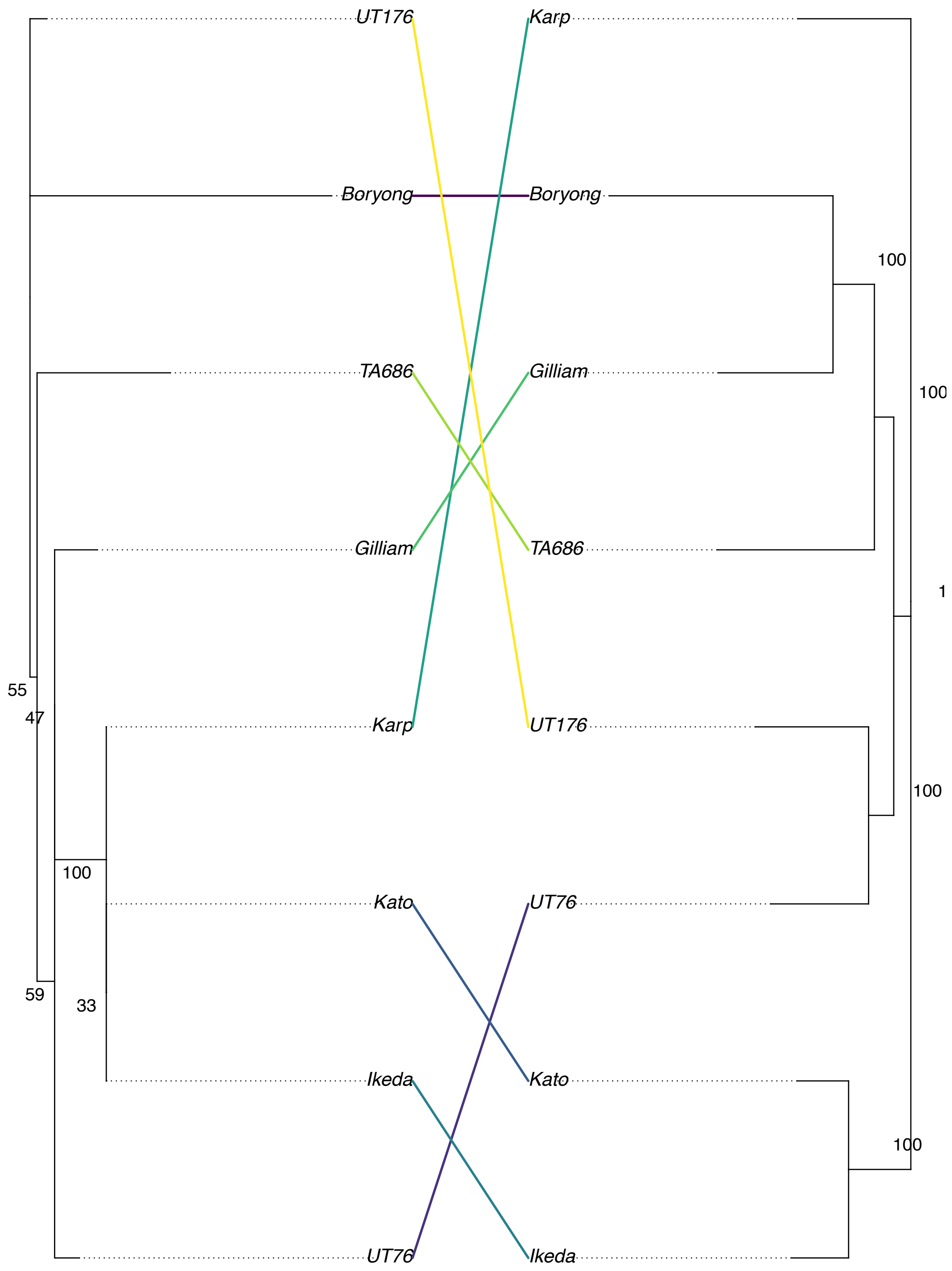
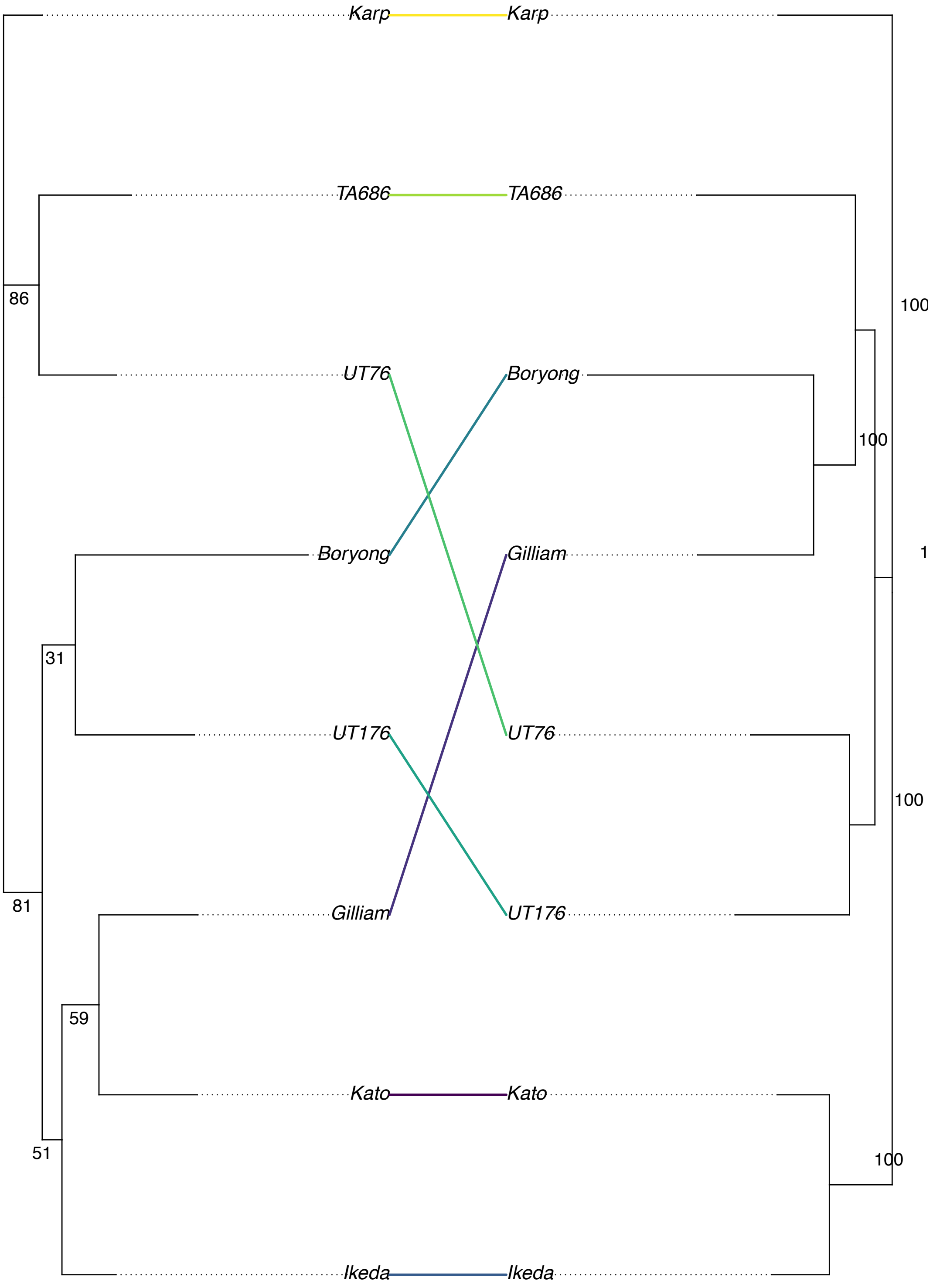


Figure S6. A phylogenetic tree showing the relationship between a tree generated using the 47kDa antigen sequences, and the sequences of 657 core genes.



679

680 *Figure S7. A phylogenetic tree showing the relationship between a tree generated using*
681 *MLST gene sequences, and the sequences of 657 core genes.*

682

Strain	Sequencing	Accession
Karp	Institute for Genome Sciences	PRJNA212440
Kato	Institute for Genome Sciences	PRJNA212441
Gilliam	Institute for Genome Sciences	PRJNA212442
TA686	MicrobesNG	PRJEB24834
TA763	Institute for Genome Sciences	PRJNA212454
FPW1038	Oxford Genomics Centre	PRJEB24834
UT76	Oxford Genomics Centre	PRJEB24834
UT176	Oxford Genomics Centre	PRJEB24834

Table S1. Sources and data accession for Illumina sequencing data.

Genome	NCBI Identifier
Orientia tsutsugamushi strain Boryong	GCF_000063545.1
Orientia tsutsugamushi strain Ikeda	GCF_000010205.1
Rickettsia typhi strain Wilmington	GCF_000008045.1
Rickettsia endosymbiont of *Ixodes scapularis*	GCF_000160735.1

Table S2. NCBI identifiers for previously published strains used in this paper.

Sample	**Genome Length**	**Length of repetitive sequence (bp)**	**Percentage of genome which is repetitive**
Boryong	2127051	895302	42
FPW1038	2035338	957348	47
Gilliam	2465012	1246424	51
Ikeda	2008987	721214	36
Karp	2469803	1210014	49
Kato	2319449	1050415	45
TA686	2254553	976333	43
TA763	2089396	895735	43
UT176	1932116	635697	33
UT76	2078193	868414	42
REIS	2100092	426115	20
Wilmington	1111496	0	0

Table S3. Total length of repetitive genome sequences in each strain, and as a percentage of the genome. REIS: Rickettsia endosymbiont of Ixodes scapularis. Wilmington: Rickettsia typhi strain Wilmington.

Sample	**Genome Length**	**Length of core genes**	**Core genes as proportion of genome**	**Length of repeat genes**	**Repeat genes as percentage of genome**
Boryong	2127051	679631	0.32	748541	35
Gilliam	2465012	681491	0.28	1165831	47
Ikeda	2008987	683889	0.34	757868	38
Karp	2469803	682061	0.28	1163785	47
Kato	2319449	682142	0.29	1039243	45
TA686	2254553	682706	0.30	933469	41
UT176	1932116	681689	0.35	738572	38
UT76	2078193	682964	0.33	826716	40

Table S4. Core gene and core repeat statistics.

Table S5. Core genes calculated by Roary. Gene names are given for the Karp strain.

Gene group	Island number	Annotation	Gene name	Boryong	Gilliam	Ikeda	Karp	Kato	TA686	UT176	UT76-HP
clpP	1	ATP-dependent Clp protease proteolytic subunit		Boryong_01567	Gilliam_01942	Ikeda_00423	Karp_01574	Kato_01535	TA686_02079	UT176_01755	UT76-HP_01648
gatB	2	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B	GatB	Boryong_01584	Gilliam_01957	Ikeda_00409	Karp_01279	Kato_01521	TA686_00288	UT176_01741	UT76-HP_01661
gatA	2	glutamyl-tRNA(Gln) amidotransferase subunit A	GatA	Boryong_01583	Gilliam_01956	Ikeda_00410	Karp_01280	Kato_01522	TA686_00289	UT176_01742	UT76-HP_01660
group_5707	2	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit C	GatC	Boryong_01582	Gilliam_01955	Ikeda_00411	Karp_01281	Kato_01523	TA686_00290	UT176_01743	UT76-HP_01659
group_6080	2	RNase J family beta-CASP ribonuclease		Boryong_01581	Gilliam_01954	Ikeda_00412	Karp_01282	Kato_01524	TA686_00291	UT176_01744	UT76-HP_01658
group_7975	2	DNA-binding response regulator		Boryong_01580	Gilliam_01953	Ikeda_00413	Karp_01283	Kato_01525	TA686_00292	UT176_01745	UT76-HP_01657
group_250	3	transposase		Boryong_00790	Gilliam_02703	Ikeda_02118	Karp_00040	Kato_00709	TA686_01162	UT176_00574	UT76-HP_00998
group_5845	3	multidrug ABC transporter ATP-binding protein		Boryong_00791	Gilliam_02704	Ikeda_02119	Karp_00041	Kato_00710	TA686_01161	UT176_00573	UT76-HP_00999
group_7831	3	UMP kinase		Boryong_00792	Gilliam_02705	Ikeda_02120	Karp_00042	Kato_00711	TA686_01160	UT176_00572	UT76-HP_01000
group_5846	3	phospho-N-acetylmuramoyl-pentapeptide- transferase	MraY	Boryong_00793	Gilliam_02706	Ikeda_02121	Karp_00043	Kato_00712	TA686_01159	UT176_00571	UT76-HP_01001
group_5550	3	UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate--D-alanyl-D-alanine ligase	MurF	Boryong_00794	Gilliam_02707	Ikeda_02122	Karp_00044	Kato_00713	TA686_01158	UT176_00570	UT76-HP_01002
group_5397	3	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diaminopimelate ligase	MurE	Boryong_00795	Gilliam_02708	Ikeda_02123	Karp_00045	Kato_00714	TA686_01157	UT176_00569	UT76-HP_01003
group_5847	3	penicillin-binding protein	PBP	Boryong_00796	Gilliam_02709	Ikeda_02124	Karp_00046	Kato_00715	TA686_01156	UT176_00568	UT76-HP_01004
ftsL	3	hypothetical protein	FtsL	Boryong_00797	Gilliam_02710	Ikeda_02125	Karp_00047	Kato_00716	TA686_01155	UT176_00567	UT76-HP_01005
group_5670	3	16S rRNA methyltransferase		Boryong_00798	Gilliam_02711	Ikeda_02126	Karp_00048	Kato_00717	TA686_01154	UT176_00566	UT76-HP_01006
group_6027	3	molecular chaperone DnaJ	DnaJ	Boryong_00799	Gilliam_02712	Ikeda_02127	Karp_00049	Kato_00718	TA686_01153	UT176_00565	UT76-HP_01007
group_7111	3	molecular chaperone DnaK	DnaK	Boryong_00800	Gilliam_02713	Ikeda_02128	Karp_00050	Kato_00719	TA686_01152	UT176_00564	UT76-HP_01008
group_6028	3	BolA family transcriptional regulator		Boryong_00801	Gilliam_02714	Ikeda_02129	Karp_00051	Kato_00720	TA686_01151	UT176_00563	UT76-HP_01009
group_5671	3	enoyl-ACP reductase	ENR	Boryong_00802	Gilliam_02715	Ikeda_02130	Karp_00052	Kato_00721	TA686_01150	UT176_00562	UT76-HP_01010
group_4752	4	sodium:proline symporter		Boryong_00980	Gilliam_00683	Ikeda_01863	Karp_00697	Kato_02375	TA686_02102	UT176_02068	UT76-HP_02241
group_5324	5	hypothetical protein		Boryong_00010	Gilliam_00014	Ikeda_01778	Karp_00009	Kato_00009	TA686_01198	UT176_00009	UT76-HP_00009
group_5345	6	hypothetical protein		Boryong_00676	Gilliam_02258	Ikeda_01084	Karp_02002	Kato_00906	TA686_02341	UT176_00940	UT76-HP_01866
group_5664	6	UDP-N-acetylmuramate--L-alanine ligase	MurC	Boryong_00675	Gilliam_02259	Ikeda_01085	Karp_02003	Kato_00905	TA686_02340	UT176_00939	UT76-HP_01867
group_5663	6	UDP-N-acetylenolpyruvoylglucosamine reductase	MurB	Boryong_00674	Gilliam_02260	Ikeda_01086	Karp_02004	Kato_00904	TA686_02339	UT176_00938	UT76-HP_01868
group_5839	6	D-alanine--D-alanine ligase	Ddl	Boryong_00673	Gilliam_02261	Ikeda_01087	Karp_02005	Kato_00903	TA686_02338	UT176_00937	UT76-HP_01869
group_5662	6	cell division protein FtsQ	FtsQ	Boryong_00672	Gilliam_02262	Ikeda_01088	Karp_02006	Kato_00902	TA686_02337	UT176_00936	UT76-HP_01870
group_5548	6	DNA replication/repair protein RecF	RecF	Boryong_00671	Gilliam_02263	Ikeda_01089	Karp_02007	Kato_00901	TA686_02336	UT176_00935	UT76-HP_01871
group_5349	7	hypothetical protein		Boryong_01099	Gilliam_00969	Ikeda_00001	Karp_02387	Kato_02127	TA686_01419	UT176_00668	UT76-HP_00144
group_6051	7	virB4 protein precursor		Boryong_01098	Gilliam_00970	Ikeda_00002	Karp_02388	Kato_02126	TA686_01418	UT176_00667	UT76-HP_00143
group_5858	7	type I glyceraldehyde-3-phosphate dehydrogenase	GapA	Boryong_01097	Gilliam_00971	Ikeda_00003	Karp_02389	Kato_02125	TA686_01417	UT176_00666	UT76-HP_00142
group_5857	7	phosphoglycerate kinase	Pgk	Boryong_01096	Gilliam_00972	Ikeda_00004	Karp_02390	Kato_02124	TA686_01416	UT176_00665	UT76-HP_00141
group_6050	7	hypothetical protein		Boryong_01095	Gilliam_00973	Ikeda_00005	Karp_02391	Kato_02123	TA686_01415	UT176_00664	UT76-HP_00140
group_5856	7	proline--tRNA ligase	ProS	Boryong_01094	Gilliam_00974	Ikeda_00006	Karp_02392	Kato_02122	TA686_01414	UT176_00663	UT76-HP_00139

group_7234	7	ATP-dependent Clp protease ATP-binding subunit ClpX	ClpX	Boryong_01093	Gilliam_00975	Ikeda_00007	Karp_02393	Kato_02121	TA686_01413	UT176_00662	UT76-HP_00138
group_8077	7	elongation factor P		Boryong_01092	Gilliam_00976	Ikeda_00008	Karp_02394	Kato_02120	TA686_01412	UT176_00661	UT76-HP_00137
group_6049	7	extragenic suppressor protein SuhB	SuhB	Boryong_01091	Gilliam_00977	Ikeda_00009	Karp_02395	Kato_02119	TA686_01411	UT176_00660	UT76-HP_00136
group_5557	7	tRNA (adenosine(37)-N6)-threonylcarbamoyltransferase complex dimerization subunit type 1 TsaB	TsaB	Boryong_01090	Gilliam_00978	Ikeda_00010	Karp_02396	Kato_02118	TA686_01410	UT176_00659	UT76-HP_00135
group_5351	8	hypothetical protein		Boryong_01511	Gilliam_01339	Ikeda_01143	Karp_01848	Kato_00841	TA686_00231	UT176_01364	UT76-HP_01169
group_5352	9	glycerol-3-phosphate dehydrogenase (NAD(P)( ))	GpsA	Boryong_01513	Gilliam_01337	Ikeda_01145	Karp_01850	Kato_00843	TA686_00229	UT176_01366	UT76-HP_01171
group_8119	9	tRNA (N(6)-L-threonylcarbamoyladenosine(37)-C(2))-methylthiotransferase MtaB	MtaB	Boryong_01514	Gilliam_01336	Ikeda_01146	Karp_01851	Kato_00844	TA686_00228	UT176_01367	UT76-HP_01172
group_5359	10	crossover junction endodeoxyribonuclease RuvC	RuvC	Boryong_01866	Gilliam_01437	Ikeda_01053	Karp_02233	Kato_00938	TA686_02092	UT176_01131	UT76-HP_01610
group_5717	10	tRNA dihydrouridine synthase DusB	DusB	Boryong_01867	Gilliam_01438	Ikeda_01054	Karp_02234	Kato_00937	TA686_02091	UT176_01132	UT76-HP_01611
group_5494	10	hypothetical protein		Boryong_01868	Gilliam_01439	Ikeda_01055	Karp_02235	Kato_00936	TA686_02090	UT176_01133	UT76-HP_01612
group_5495	10	bifunctional 3-demethylubiquinone 3-O-methyltransferase/2-octaprenyl-6-hydroxy phenol methylase		Boryong_01869	Gilliam_01440	Ikeda_01056	Karp_02236	Kato_00935	TA686_02089	UT176_01134	UT76-HP_01613
group_5718	10	protein-(glutamine-N5) methyltransferase, release factor-specific		Boryong_01870	Gilliam_01441	Ikeda_01057	Karp_02237	Kato_00934	TA686_02088	UT176_01135	UT76-HP_01614
group_6104	10	tRNA pseudouridine(38-40) synthase TruA	TruA	Boryong_01871	Gilliam_01442	Ikeda_01058	Karp_02238	Kato_00933	TA686_02087	UT176_01136	UT76-HP_01615
group_7746	10	50S ribosomal protein L13	L13	Boryong_01872	Gilliam_01443	Ikeda_01059	Karp_02239	Kato_00932	TA686_02086	UT176_01137	UT76-HP_01616
group_5719	10	30S ribosomal protein S9	S9	Boryong_01873	Gilliam_01444	Ikeda_01060	Karp_02240	Kato_00931	TA686_02085	UT176_01138	UT76-HP_01617
group_5458	11	rRNA (cytidine-2'-O-)-methyltransferase		Boryong_01202	Gilliam_01733	Ikeda_00482	Karp_01858	Kato_01661	TA686_00545	UT176_01357	UT76-HP_01640
group_5865	11	serine--tRNA ligase	SerS	Boryong_01203	Gilliam_01734	Ikeda_00481	Karp_01859	Kato_01662	TA686_00546	UT176_01356	UT76-HP_01639
group_7705	11	twin-arginine translocase subunit TatC	TatC	Boryong_01204	Gilliam_01735	Ikeda_00480	Karp_01860	Kato_01663	TA686_00547	UT176_01355	UT76-HP_01638
group_6566	11	hypothetical protein		Boryong_01205	Gilliam_01736	Ikeda_00479	Karp_01861	Kato_01664	TA686_00548	UT176_01354	UT76-HP_01637
group_6058	11	16S rRNA methyltransferase		Boryong_01206	Gilliam_01737	Ikeda_00478	Karp_01862	Kato_01665	TA686_00549	UT176_01353	UT76-HP_01636
group_7851	11	chromosome partitioning protein ParA	ParA	Boryong_01207	Gilliam_01738	Ikeda_00477	Karp_01863	Kato_01666	TA686_00550	UT176_01352	UT76-HP_01635
group_6059	11	chromosome partitioning protein	ParB	Boryong_01208	Gilliam_01739	Ikeda_00476	Karp_01864	Kato_01667	TA686_00551	UT176_01351	UT76-HP_01634
group_5485	12	rod shape-determining protein MreC	MreC	Boryong_01561	Gilliam_02176	Ikeda_00336	Karp_01810	Kato_01222	TA686_01169	UT176_01839	UT76-HP_01300
group_7287	12	rod shape-determining protein	MreB	Boryong_01562	Gilliam_02175	Ikeda_00335	Karp_01811	Kato_01223	TA686_01168	UT176_01838	UT76-HP_01301
group_5575	12	dihydrolipoamide acetyltransferase		Boryong_01563	Gilliam_02174	Ikeda_00334	Karp_01812	Kato_01224	TA686_01167	UT176_01837	UT76-HP_01302
group_5491	13	aspartate kinase	AK	Boryong_01771	Gilliam_01906	Ikeda_00772	Karp_01348	Kato_01421	TA686_00345	UT176_01725	UT76-HP_01353
group_5712	13	hypothetical protein		Boryong_01772	Gilliam_01905	Ikeda_00773	Karp_01349	Kato_01420	TA686_00346	UT176_01724	UT76-HP_01354
group_8049	13	potassium transporter		Boryong_01773	Gilliam_01904	Ikeda_00774	Karp_01350	Kato_01419	TA686_00347	UT176_01723	UT76-HP_01355
group_5713	13	5-formyltetrahydrofolate cyclo-ligase	YgfA	Boryong_01774	Gilliam_01903	Ikeda_00775	Karp_01351	Kato_01418	TA686_00348	UT176_01722	UT76-HP_01356
group_5579	13	hypothetical protein		Boryong_01775	Gilliam_01902	Ikeda_00776	Karp_01352	Kato_01417	TA686_00349	UT176_01721	UT76-HP_01357
group_5496	14	ankyrin repeat-containing protein 13	Ank13	Boryong_01925	Gilliam_01541	Ikeda_00523	Karp_01630	Kato_01772	TA686_00095	UT176_01272	UT76-HP_01784
group_5411	14	hypothetical protein		Boryong_01926	Gilliam_01542	Ikeda_00522	Karp_01631	Kato_01773	TA686_00094	UT176_01273	UT76-HP_01785
group_5497	15	heme A synthase		Boryong_02136	Gilliam_01573	Ikeda_00787	Karp_01402	Kato_01407	TA686_01213	UT176_01711	UT76-HP_01347

group_5549	16	threonylcarbamoyl-AMP synthase	TsaC	Boryong_00680	Gilliam_02254	Ikeda_01080	Karp_01998	Kato_00910	TA686_02345	UT176_00944	UT76-HP_01862
group_5448	16	glycine--tRNA ligase subunit beta	GlyS	Boryong_00679	Gilliam_02255	Ikeda_01081	Karp_01999	Kato_00909	TA686_02344	UT176_00943	UT76-HP_01863
group_5840	16	glycine--tRNA ligase subunit alpha	GlyQ	Boryong_00678	Gilliam_02256	Ikeda_01082	Karp_02000	Kato_00908	TA686_02343	UT176_00942	UT76-HP_01864
group_5574	17	competence protein ComEC	ComEC	Boryong_01456	Gilliam_02183	Ikeda_00344	Karp_01804	Kato_01216	TA686_00950	UT176_01848	UT76-HP_01293
group_5585	18	hypothetical protein		Boryong_01875	Gilliam_01446	Ikeda_01062	Karp_02242	Kato_00929	TA686_02083	UT176_01140	UT76-HP_01619
group_5622	19	hypothetical protein		Boryong_00133	Gilliam_01746	Ikeda_01832	Karp_00722	Kato_02398	TA686_01802	UT176_02092	UT76-HP_02218
group_5776	19	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	DapD	Boryong_00132	Gilliam_01745	Ikeda_01831	Karp_00723	Kato_02399	TA686_01803	UT176_02093	UT76-HP_02217
group_5656	20	MFS transporter permease		Boryong_00574	Gilliam_00985	Ikeda_01615	Karp_00351	Kato_00491	TA686_00383	UT176_00185	UT76-HP_00366
group_5544	20	sodium:pantothenate symporter		Boryong_00573	Gilliam_00986	Ikeda_01616	Karp_00352	Kato_00492	TA686_00382	UT176_00186	UT76-HP_00367
group_5684	21	SAM-dependent methyltransferase		Boryong_00940	Gilliam_02522	Ikeda_01914	Karp_00101	Kato_02345	TA686_00107	UT176_00553	UT76-HP_01014
group_5705	22	two-component sensor histidine kinase		Boryong_01454	Gilliam_02181	Ikeda_00342	Karp_01806	Kato_01218	TA686_00952	UT176_01846	UT76-HP_01295
group_6160	22	sigma-54-dependent Fis family transcriptional regulator		Boryong_01453	Gilliam_02180	Ikeda_00341	Karp_01807	Kato_01219	TA686_00953	UT176_01845	UT76-HP_01296
group_5483	22	hypothetical protein		Boryong_01452	Gilliam_02179	Ikeda_00340	Karp_01808	Kato_01220	TA686_00954	UT176_01844	UT76-HP_01297
group_5722	23	aspartate aminotransferase	AspC	Boryong_02006	Gilliam_00094	Ikeda_01340	Karp_00229	Kato_00160	TA686_02116	UT176_00101	UT76-HP_00570
ubiG	23	Ubiquinone biosynthesis O-methyltransferase	UbiG	Boryong_02007	Gilliam_00095	Ikeda_01341	Karp_00230	Kato_00161	TA686_02115	UT176_00102	UT76-HP_00569
group_5723	23	ABC transporter		Boryong_02008	Gilliam_00096	Ikeda_01342	Karp_00231	Kato_00162	TA686_02114	UT176_00103	UT76-HP_00568
group_5724	23	hypothetical protein		Boryong_02009	Gilliam_00097	Ikeda_01343	Karp_00232	Kato_00163	TA686_02113	UT176_00104	UT76-HP_00567
group_5900	23	coproporphyrinogen III oxidase		Boryong_02010	Gilliam_00098	Ikeda_01344	Karp_00233	Kato_00164	TA686_02112	UT176_00105	UT76-HP_00566
group_6112	23	hypothetical protein		Boryong_02011	Gilliam_00099	Ikeda_01345	Karp_00234	Kato_00165	TA686_02111	UT176_00106	UT76-HP_00565
group_5587	23	DNA repair protein RecO	RecO	Boryong_02012	Gilliam_00100	Ikeda_01346	Karp_00235	Kato_00166	TA686_02110	UT176_00107	UT76-HP_00564
group_5732	24	DNA helicase II	UvrD	Boryong_02217	Gilliam_00618	Ikeda_00573	Karp_01153	Kato_01479	TA686_01997	UT176_01493	UT76-HP_01067
group_5740	25	NAD-glutamate dehydrogenase	GdhA	Boryong_02452	Gilliam_01083	Ikeda_02116	Karp_02529	Kato_00707	TA686_01765	UT176_00639	UT76-HP_00743
group_5739	25	tRNA uridine-5-carboxymethylaminomethyl(34) synthesis GTPase MnmE		Boryong_02451	Gilliam_01084	Ikeda_02115	Karp_02530	Kato_00706	TA686_01766	UT176_00640	UT76-HP_00744
group_6137	25	recombinase XerC	XerC	Boryong_02450	Gilliam_01085	Ikeda_02114	Karp_02531	Kato_00705	TA686_01767	UT176_00641	UT76-HP_00745
group_6136	25	RNA polymerase-binding protein DksA	DksA	Boryong_02449	Gilliam_01086	Ikeda_02113	Karp_02532	Kato_00704	TA686_01768	UT176_00642	UT76-HP_00746
group_6135	25	inorganic pyrophosphatase		Boryong_02448	Gilliam_01087	Ikeda_02112	Karp_02533	Kato_00703	TA686_01769	UT176_00643	UT76-HP_00747
group_5415	25	DNA polymerase III subunit delta'	HolB	Boryong_02447	Gilliam_01088	Ikeda_02111	Karp_02534	Kato_00702	TA686_01770	UT176_00644	UT76-HP_00748
group_5922	25	ribosomal large subunit pseudouridine synthase		Boryong_02446	Gilliam_01089	Ikeda_02110	Karp_02535	Kato_00701	TA686_01771	UT176_00645	UT76-HP_00749
group_5791	26	tetraacyldisaccharide 4'-kinase	IpxK	Boryong_00218	Gilliam_01219	Ikeda_02020	Karp_00613	Kato_00657	TA686_00253	UT176_00427	UT76-HP_02120
group_7640	26	hypothetical protein		Boryong_00219	Gilliam_01220	Ikeda_02021	Karp_00614	Kato_00658	TA686_00254	UT176_00428	UT76-HP_02119
group_5849	27	transporter		Boryong_00935	Gilliam_02519	Ikeda_01910	Karp_00104	Kato_02349	TA686_01643	UT176_00549	UT76-HP_01018
glpE	27	hypothetical protein		Boryong_00934	Gilliam_02518	Ikeda_01909	Karp_00105	Kato_02350	TA686_01642	UT176_00548	UT76-HP_01019
group_5864	28	protein translocase subunit SecF	SecF	Boryong_01198	Gilliam_01729	Ikeda_00486	Karp_01854	Kato_01657	TA686_00541	UT176_01361	UT76-HP_01644
group_6056	28	DNA mismatch repair protein MutS	MutS	Boryong_01199	Gilliam_01730	Ikeda_00485	Karp_01855	Kato_01658	TA686_00542	UT176_01360	UT76-HP_01643

group_6057	28	ATP/ADP translocase		Boryong_01200	Gilliam_01731	Ikeda_00484	Karp_01856	Kato_01659	TA686_00543	UT176_01359	UT76-HP_01642
group_5882	29	haloacid dehalogenase		Boryong_01633	Gilliam_01659	Ikeda_00944	Karp_01021	Kato_01991	TA686_00732	UT176_01408	UT76-HP_00989
group_7708	29	DNA gyrase subunit B	GyrB	Boryong_01632	Gilliam_01658	Ikeda_00945	Karp_01022	Kato_01992	TA686_00733	UT176_01407	UT76-HP_00988
group_5577	29	hypothetical protein		Boryong_01631	Gilliam_01657	Ikeda_00946	Karp_01023	Kato_01993	TA686_00734	UT176_01406	UT76-HP_00987
group_5881	29	amino acid permease		Boryong_01630	Gilliam_01656	Ikeda_00947	Karp_01024	Kato_01994	TA686_00735	UT176_01405	UT76-HP_00986
group_5895	30	succinate dehydrogenase iron-sulfur subunit	SdhB	Boryong_01938	Gilliam_00945	Ikeda_02208	Karp_02372	Kato_02140	TA686_00235	UT176_00689	UT76-HP_00157
group_6106	30	succinate dehydrogenase flavoprotein subunit	SdhA	Boryong_01937	Gilliam_00944	Ikeda_02209	Karp_02373	Kato_02139	TA686_00236	UT176_00688	UT76-HP_00156
group_5586	30	succinate dehydrogenase, hydrophobic membrane anchor protein	SdhD	Boryong_01936	Gilliam_00943	Ikeda_02210	Karp_02374	Kato_02138	TA686_00237	UT176_00687	UT76-HP_00155
group_5721	30	succinate dehydrogenase, cytochrome b556 subunit	SdhC	Boryong_01935	Gilliam_00942	Ikeda_02211	Karp_02375	Kato_02137	TA686_00238	UT176_00686	UT76-HP_00154
group_5901	31	hypothetical protein		Boryong_02018	Gilliam_00108	Ikeda_01353	Karp_00239	Kato_00172	TA686_01204	UT176_00113	UT76-HP_00558
group_6113	31	hypothetical protein		Boryong_02019	Gilliam_00109	Ikeda_01354	Karp_00240	Kato_00173	TA686_01205	UT176_00114	UT76-HP_00557
group_8000	31	30S ribosomal protein S12	RpsL	Boryong_02020	Gilliam_00110	Ikeda_01355	Karp_00241	Kato_00174	TA686_01206	UT176_00115	UT76-HP_00556
group_7893	31	30S ribosomal protein S7	RpsG	Boryong_02021	Gilliam_00111	Ikeda_01356	Karp_00242	Kato_00175	TA686_01207	UT176_00116	UT76-HP_00555
group_7759	31	elongation factor G	EfG	Boryong_02022	Gilliam_00112	Ikeda_01357	Karp_00243	Kato_00176	TA686_01208	UT176_00117	UT76-HP_00554
group_5902	31	30S ribosomal protein S1	RpsA	Boryong_02023	Gilliam_00113	Ikeda_01358	Karp_00244	Kato_00177	TA686_01209	UT176_00118	UT76-HP_00553
group_6017	32	50S ribosomal protein L20	RplT	Boryong_00628	Gilliam_02347	Ikeda_00902	Karp_00926	Kato_01948	TA686_00833	UT176_01165	UT76-HP_00870
group_7830	32	50S ribosomal protein L35	RpmL	Boryong_00627	Gilliam_02348	Ikeda_00903	Karp_00927	Kato_01949	TA686_00834	UT176_01166	UT76-HP_00871
group_8150	32	molecular chaperone HtpG	HtpG	Boryong_00626	Gilliam_02349	Ikeda_00904	Karp_00928	Kato_01950	TA686_00835	UT176_01167	UT76-HP_00872
group_5657	32	succinyl-diaminopimelate desuccinylase	DapE	Boryong_00625	Gilliam_02350	Ikeda_00905	Karp_00929	Kato_01951	TA686_00836	UT176_01168	UT76-HP_00873
group_6033	33	DNA translocase FtsK	FtsK	Boryong_00894	Gilliam_01256	Ikeda_00430	Karp_01269	Kato_01542	TA686_00583	UT176_01646	UT76-HP_01670
group_5682	33	hypothetical protein		Boryong_00895	Gilliam_01257	Ikeda_00429	Karp_01270	Kato_01541	TA686_00584	UT176_01647	UT76-HP_01669
group_7304	33	energy-dependent translational throttle protein EttA	EttA	Boryong_00896	Gilliam_01258	Ikeda_00428	Karp_01271	Kato_01540	TA686_00585	UT176_01648	UT76-HP_01668
group_5348	33	hypothetical protein		Boryong_00897	Gilliam_01259	Ikeda_00427	Karp_01272	Kato_01539	TA686_00586	UT176_01649	UT76-HP_01667
group_6064	34	alpha/beta hydrolase		Boryong_01286	Gilliam_01621	Ikeda_00655	Karp_01105	Kato_01727	TA686_00811	UT176_01536	UT76-HP_00785
group_5701	34	iron-sulfur-binding protein		Boryong_01285	Gilliam_01620	Ikeda_00654	Karp_01106	Kato_01726	TA686_00812	UT176_01535	UT76-HP_00784
group_5400	34	aminotransferase class V-fold PLP-dependent enzyme		Boryong_01284	Gilliam_01619	Ikeda_00653	Karp_01107	Kato_01725	TA686_00813	UT176_01534	UT76-HP_00783
group_6063	34	cysteine desulfurase		Boryong_01283	Gilliam_01618	Ikeda_00652	Karp_01108	Kato_01724	TA686_00814	UT176_01533	UT76-HP_00782
group_5868	34	iron-sulfur cluster scaffold-like protein		Boryong_01282	Gilliam_01617	Ikeda_00651	Karp_01109	Kato_01723	TA686_00815	UT176_01532	UT76-HP_00781
group_6062	34	iron-sulfur cluster assembly accessory protein		Boryong_01281	Gilliam_01616	Ikeda_00650	Karp_01110	Kato_01722	TA686_00816	UT176_01531	UT76-HP_00780
group_5564	34	co-chaperone HscB	HscB	Boryong_01280	Gilliam_01615	Ikeda_00649	Karp_01111	Kato_01721	TA686_00817	UT176_01530	UT76-HP_00779
group_5867	34	molecular chaperone HscA	HscA	Boryong_01279	Gilliam_01614	Ikeda_00648	Karp_01112	Kato_01720	TA686_00818	UT176_01529	UT76-HP_00778
group_5563	34	(2Fe-2S) ferredoxin		Boryong_01278	Gilliam_01613	Ikeda_00647	Karp_01113	Kato_01719	TA686_00819	UT176_01528	UT76-HP_00777
group_6072	35	electron transporter		Boryong_01395	Gilliam_00196	Ikeda_01076	Karp_01763	Kato_00913	TA686_02251	UT176_01154	UT76-HP_01631

group_6074	36	single-stranded DNA-binding protein		Boryong_01420	Gilliam_01983	Ikeda_00830	Karp_01239	Kato_01364	TA686_01786	UT176_01802	UT76-HP_01743
group_5704	36	hypothetical protein		Boryong_01419	Gilliam_01982	Ikeda_00831	Karp_01240	Kato_01363	TA686_01785	UT176_01803	UT76-HP_01742
group_6078	37	malate dehydrogenase	Mdh	Boryong_01520	Gilliam_01526	Ikeda_00663	Karp_01409	Kato_01734	TA686_02541	UT176_01450	UT76-HP_01425
group_6077	37	permease		Boryong_01519	Gilliam_01527	Ikeda_00664	Karp_01410	Kato_01735	TA686_02540	UT176_01451	UT76-HP_01426
group_5484	37	hypothetical protein		Boryong_01518	Gilliam_01528	Ikeda_00665	Karp_01411	Kato_01736	TA686_02539	UT176_01452	UT76-HP_01427
group_6083	38	cytochrome b	CybB	Boryong_01614	Gilliam_02691	Ikeda_00819	Karp_01467	Kato_01374	TA686_00793	UT176_01660	UT76-HP_01121
group_5878	38	ubiquinol-cytochrome c reductase iron-sulfur subunit	PetA	Boryong_01613	Gilliam_02690	Ikeda_00820	Karp_01468	Kato_01373	TA686_00792	UT176_01661	UT76-HP_01120
group_5486	38	hypothetical protein		Boryong_01612	Gilliam_02689	Ikeda_00821	Karp_01469	Kato_01372	TA686_00791	UT176_01662	UT76-HP_01119
group_5877	38	heme exporter protein B	CcmB	Boryong_01611	Gilliam_02688	Ikeda_00822	Karp_01470	Kato_01371	TA686_00790	UT176_01663	UT76-HP_01118
group_5709	38	cytochrome c biogenesis protein CcmA	CcmA	Boryong_01610	Gilliam_02687	Ikeda_00823	Karp_01471	Kato_01370	TA686_00789	UT176_01664	UT76-HP_01117
group_6087	39	2-hydroxyacid dehydrogenase		Boryong_01640	Gilliam_01666	Ikeda_00937	Karp_01014	Kato_01984	TA686_00725	UT176_01415	UT76-HP_00996
group_7914	39	cation:proton antiporter		Boryong_01639	Gilliam_01665	Ikeda_00938	Karp_01015	Kato_01985	TA686_00726	UT176_01414	UT76-HP_00995
group_6086	39	cation:proton antiporter		Boryong_01638	Gilliam_01664	Ikeda_00939	Karp_01016	Kato_01986	TA686_00727	UT176_01413	UT76-HP_00994
group_5883	39	sodium:proton antiporter		Boryong_01637	Gilliam_01663	Ikeda_00940	Karp_01017	Kato_01987	TA686_00728	UT176_01412	UT76-HP_00993
group_5710	39	sodium:proton antiporter		Boryong_01636	Gilliam_01662	Ikeda_00941	Karp_01018	Kato_01988	TA686_00729	UT176_01411	UT76-HP_00992
group_8081	39	sodium:proton antiporter		Boryong_01635	Gilliam_01661	Ikeda_00942	Karp_01019	Kato_01989	TA686_00730	UT176_01410	UT76-HP_00991
group_6098	40	hypothetical protein		Boryong_01851	Gilliam_01422	Ikeda_01037	Karp_02217	Kato_00953	TA686_01496	UT176_01371	UT76-HP_01595
group_6107	41	S26 family signal peptidase		Boryong_01941	Gilliam_00952	Ikeda_02205	Karp_00806	Kato_02077	TA686_00272	UT176_00890	UT76-HP_02017
group_6108	41	ribonuclease III	Rnc	Boryong_01942	Gilliam_00953	Ikeda_02204	Karp_00807	Kato_02076	TA686_00273	UT176_00889	UT76-HP_02016
group_6121	42	nucleoside-diphosphate kinase	Ndk	Boryong_02109	Gilliam_00855	Ikeda_02149	Karp_02170	Kato_00743	TA686_01081	UT176_01940	UT76-HP_00186
group_6122	43	hypothetical protein		Boryong_02132	Gilliam_01567	Ikeda_00791	Karp_01397	Kato_01402	TA686_02328	UT176_01708	UT76-HP_01343
group_6132	44	phospholipase D family protein		Boryong_02222	Gilliam_00612	Ikeda_00567	Karp_01146	Kato_01485	TA686_01628	UT176_01487	UT76-HP_01062
group_6754	45	elongation factor 4	IepA	Boryong_01410	Gilliam_02108	Ikeda_00834	Karp_01242	Kato_01360	TA686_01368	UT176_01301	UT76-HP_01739
group_5874	45	peptide chain release factor 1	PrfA	Boryong_01409	Gilliam_02107	Ikeda_00835	Karp_01243	Kato_01359	TA686_01367	UT176_01302	UT76-HP_01738
group_7286	46	DNA-binding protein		Boryong_00490	Gilliam_00721	Ikeda_00027	Karp_02417	Kato_02102	TA686_00757	UT176_00621	UT76-HP_01986
surA	46	Chaperone SurA	SurA	Boryong_00489	Gilliam_00720	Ikeda_00028	Karp_02418	Kato_02101	TA686_00758	UT176_00620	UT76-HP_01985
group_6007	46	16S rRNA (adenine(1518)-N(6)/adenine(1519)-N(6))-dimethyltransferase		Boryong_00488	Gilliam_00719	Ikeda_00029	Karp_02419	Kato_02100	TA686_00759	UT176_00619	UT76-HP_01984
group_5824	46	DNA recombination protein RmuC	RmuC	Boryong_00487	Gilliam_00718	Ikeda_00030	Karp_02420	Kato_02099	TA686_00760	UT176_00618	UT76-HP_01983
group_5650	46	zinc metalloprotease		Boryong_00486	Gilliam_00717	Ikeda_00031	Karp_02421	Kato_02098	TA686_00761	UT176_00617	UT76-HP_01982
group_5539	46	outer membrane protein assembly factor BamA	BamA	Boryong_00485	Gilliam_00716	Ikeda_00032	Karp_02422	Kato_02097	TA686_00762	UT176_00616	UT76-HP_01981
group_5649	46	thiol reductase thioredoxin		Boryong_00484	Gilliam_00715	Ikeda_00033	Karp_02423	Kato_02096	TA686_00763	UT176_00615	UT76-HP_01980
group_7769	47	thioredoxin-disulfide reductase	TrxB	Boryong_00020	Gilliam_00024	Ikeda_01757	Karp_00011	Kato_00071	TA686_02375	UT176_00364	UT76-HP_00026
group_7112	47	permease		Boryong_00021	Gilliam_00025	Ikeda_01758	Karp_00012	Kato_00070	TA686_02376	UT176_00363	UT76-HP_00025
group_5621	47	translocation protein TolB	TolB	Boryong_00022	Gilliam_00026	Ikeda_01759	Karp_00013	Kato_00069	TA686_02377	UT176_00362	UT76-HP_00024

group_5775	47	dihydrolipoyl dehydrogenase	lpdA	Boryong_00023	Gilliam_00027	Ikeda_01760	Karp_00014	Kato_00068	TA686_02378	UT176_00361	UT76-HP_00023
group_5425	47	SAM-dependent methyltransferase		Boryong_00024	Gilliam_00028	Ikeda_01761	Karp_00015	Kato_00067	TA686_02379	UT176_00360	UT76-HP_00022
group_5426	47	hypothetical protein		Boryong_00025	Gilliam_00029	Ikeda_01762	Karp_00016	Kato_00066	TA686_02380	UT176_00359	UT76-HP_00021
group_7894	48	type I methionyl aminopeptidase	Map	Boryong_01573	Gilliam_01947	Ikeda_00420	Karp_01288	Kato_01532	TA686_00765	UT176_01751	UT76-HP_01652
group_7905	49	ubiquinone biosynthesis protein UbiB	UbiB	Boryong_01795	Gilliam_02198	Ikeda_00705	Karp_01872	Kato_01873	TA686_00284	UT176_01700	UT76-HP_01485
group_5580	49	ubiquinone biosynthesis protein	UbiJ	Boryong_01796	Gilliam_02197	Ikeda_00704	Karp_01873	Kato_01872	TA686_00283	UT176_01701	UT76-HP_01486
group_6093	49	ribosome maturation factor		Boryong_01797	Gilliam_02196	Ikeda_00703	Karp_01874	Kato_01871	TA686_00282	UT176_01702	UT76-HP_01487
group_6094	49	transcription termination/antitermination protein NusA	NusA	Boryong_01798	Gilliam_02195	Ikeda_00702	Karp_01875	Kato_01870	TA686_00281	UT176_01703	UT76-HP_01488
group_5889	49	translation initiation factor IF-2	InfB	Boryong_01799	Gilliam_02194	Ikeda_00701	Karp_01876	Kato_01869	TA686_00280	UT176_01704	UT76-HP_01489
group_7895	49	ribosome-binding factor A	RbfA	Boryong_01800	Gilliam_02193	Ikeda_00700	Karp_01877	Kato_01868	TA686_00279	UT176_01705	UT76-HP_01490
group_7960	50	preprotein translocase subunit YajC	YajC	Boryong_02000	Gilliam_00087	Ikeda_01330	Karp_00223	Kato_00152	TA686_02313	UT176_00845	UT76-HP_00576
group_6110	50	protein translocase subunit SecD	SecD	Boryong_02001	Gilliam_00088	Ikeda_01331	Karp_00224	Kato_00153	TA686_02312	UT176_00846	UT76-HP_00575
group_8117	51	peptidase S66		Boryong_00304	Gilliam_00580	Ikeda_01592	Karp_00511	Kato_00468	TA686_01761	UT176_00261	UT76-HP_00436

Product		Boryong	Gilliam	Ikeda	Karp	Kato	TA686	UT176	UT76
(p)pGpp hydrolase		37	31	25	40	26	14	16	25
(p)ppGpp synthetase		2	2	1	5	1	0	2	2
spoT ppGpp hydrolase		3	15	7	16	9	11	5	5
ABC transporter ATP-binding protein		1	2	2	2	3	2	1	3
Aconitate hydratase A		1	1	2	1	2	0	1	1
All ankyrin proteins		43	46	40	58	37	39	37	38
	ankyrin	14	26	18	33	23	21	21	25
	ankyrin repeat-containing protein	13	10	13	11	7	4	10	8
	ankyrin repeat-containing protein 09	4	6	3	4	2	8	3	3
	ankyrin repeat-containing protein 13	3	1	1	1	1	0	1	1
	ankyrin repeat-containing protein 16	9	0	2	7	2	2	1	0
	ankyrin repeat-containing protein 17	0	1	1	1	0	4	1	0
	ankyrin repeat-containing protein 19	0	2	2	1	2	0	0	1
ATP-binding protein		48	85	63	99	91	97	44	87
Cell division protein FtsB		1	2	1	1	1	1	1	1
All conjugal transfer proteins		461	532	378	570	502	462	330	481
	conjugal transfer protein	166	202	138	242	181	202	137	194
	conjugal transfer protein TraA	75	86	60	83	56	64	41	62
	conjugal transfer protein TraC	70	50	39	40	65	37	34	61
	conjugal transfer protein TraD	1	0	2	2	2	2	0	0
	conjugal transfer protein TraG	13	29	21	28	24	25	19	24

	conjugal transfer protein TraH	41	37	37	44	52	34	24	41
	conjugal transfer protein TraI	41	62	32	65	48	50	25	33
	conjugal transfer protein TraN	46	49	30	38	40	27	36	45
	type-F conjugative transfer system pilin assembly protein TrbC	0	0	0	2	0	0	0	0
	type-F conjugative transfer system protein TraW	8	17	19	26	34	21	14	21
deoxyribodipyrimidine photo-lyase		4	1	1	1	1	0	1	0
DNA helicase		0	0	1	3	6	0	0	1
DNA methyltransferase		27	32	17	29	26	22	17	28
DNA polymerase III subunit epsilon		1	1	1	1	1	2	1	1
elongation factor Tu		2	2	2	2	2	2	2	2
exodeoxyribonuclease III		3	1	4	4	2	1	2	3
exodeoxyribonuclease VII small subunit		1	1	1	1	1	2	1	1
Group II intron-encoded protein LtrA		0	0	0	0	0	4	0	0
guanosine polyphosphate pyrophosphohydrolase		3	2	10	3	9	11	1	5
helix-turn-helix domain-containing protein		3	0	0	0	0	4	0	0
histidine kinase		1	8	9	8	13	16	1	9
HNH endonuclease		4	2	1	32	19	37	0	3
hydrolase		5	13	13	11	20	12	7	14
hypothetical protein		321	250	180	259	241	242	134	188
integrase		69	77	69	71	92	87	44	82

All transposases		338	602	306	325	242	409	487	242
	DDE transposase family protein	0	0	4	2	1	3	0	1
	IS110 family transposase	19	8	34	22	14	23	13	5
	IS5 family transposase ISOt6	199	157	101	143	85	163	73	87
	IS630 family transposase	26	342	71	27	37	83	316	29
	transposase	94	95	96	131	105	137	85	120
lipase LipB		1	1	1	1	1	1	0	2
lysine--tRNA ligase		1	1	1	1	1	2	1	1
membrane protein		12	27	17	34	25	20	16	22
N-6 DNA methylase		6	1	0	0	0	0	0	0
NADP-dependent oxidoreductase		1	1	2	1	1	1	1	1
peroxiredoxin		1	2	4	4	7	5	2	2
phosphatidate cytidylyltransferase		1	2	1	1	1	1	1	1
phosphoribosylaminoimidazolesuccinocarboxamide synthase		1	1	1	1	1	3	2	2
polyribonucleotide nucleotidyltransferase		1	1	1	1	1	1	1	2
preprotein translocase SecA subunit-like protein		0	4	2	7	2	9	0	2
Propionyl-CoA carboxylase beta chain		0	1	2	1	2	0	1	1
repeat-containing protein D		4	0	4	1	2	0	1	1
replicative DNA helicase		47	33	28	40	36	39	17	34
reverse transcriptase		58	19	32	5	33	23	2	6
RNA-binding protein		3	2	5	10	3	12	4	4

sodium:proline symporter		4	4	7	6	8	5	5	5
TAL effector protein PthXo1		1	0	3	2	3	0	1	3
All TPR repeat-containing proteins		22	40	18	29	37	24	22	27
	TPR repeat-containing protein 03	0	12	6	8	10	7	11	4
	TPR repeat-containing protein 08	22	28	12	21	27	17	11	23
tryptophan--tRNA ligase		2	1	1	1	1	1	1	1
UDP pyrophosphate synthase		1	2	1	1	1	1	1	1

Table S6. Repeat gene counts in each strain. Repeat genes were grouped by protein similarity and annotated with the product of the longest gene in the group where annotations differed.

Sample	Pseudogenes	Truncated 5'	Truncated 3'	Frameshift
Boryong	432	219	302	46
Gilliam	484	262	278	51
Ikeda	257	141	186	38
Karp	321	105	236	47
Kato	286	143	178	57
TA686	453	200	307	50
UT176	465	107	392	53
UT76	319	149	203	52

Table S7. Pseudogenes and causes of pseudogenisation for each strain. The causes are not mutually exclusive, and may sum to greater than the total number of pseudogenes.

	56kDa	47kDa	MLST	Core genome
56kDa	-	10	10	8
47kDa	10	-	8	6
MLST	10	8	-	10
Core genome	8	6	10	-

715 *Table S8. Robinson-Foulds distances between phylogenetic trees.*