Worldwide Population Structure of *Escherichia coli* Reveals Two Major Subspecies

Yu Kang[1†], Lina Yuan1[1†‡], Zilong He[1†], Fei Chen[1†], Zhancheng Gao[2†], Shulin Liu[3,4], Xinmiao Jia[1], Qin Ma[5], Xinhao Jin[1], Rongrong Fu[1], Yang Yu[6,7], Chunxiong Luo[8], Jiayan Wu[1], Jingfa Xiao[1], Songnian Hu[1], Jun Yu[1]*.

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, 100101, Beijing, PR China.

[2] Department of Respiratory & Critical Care Medicine, Peking University People's Hospital, Beijing, 100044, PR China

[3] Genomics Research Center (one of The State-Province Key Laboratories of Biomedicine-Pharmaceutics of China), Harbin Medical University, Harbin, 150081, PR China

[4] Department of Microbiology and Infectious Diseases, University of Calgary, Alberta, Canada, T2N 1N4.

[5] Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, Brookings, SD, 57007, USA.

[6] School of Life Sciences, Liaoning University, Shenyang, 110036, PR China

[7] Department of Biology, University of Virginia, Charlottesville, Virginia, 22904, USA

[8] Center for Microfluidic and Nanotechnology, State Key Laboratory for Artificial Microstructures and Mesoscopic Physics, School of Physics, Peking University, Beijing, 100871, PR China

* To whom correspondence should be addressed. CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, P.R. China. Tel: +8610-84097898; Fax: +8610-84097540; Email: junyu@big.ac.cn

†The first 5 authors should be regarded as joint First Authors.

‡Current Address: Lina Yuan, Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing100005; Suzhou Institute of Systems Medicine, Suzhou 215123, P.R.China.

Running Head: Two Subspecies of *Escherichia coli*

Word Count:

Abstract 171

Text 5742

ABSTRACT

Escherichia coli is a Gram-negative bacterial species with both great biological diversity and important clinical relevance. To study its population structure in both world-wide and genome-wide scales, we scrutinise phylogenetically 104 high-quality complete genomes of diverse human/animal hosts, among which 45 are new additions to the collection; most of them are clustered into two major clades: Vig (vigorous) and Slu (sluggish). The two clades not only show distinct physiological features but also genome content and sequence variation. Limited recombination and horizontal gene transfer separate the two clades, as opposed to extensive intra-clade gene flow that functionally homogenizes even commensal and pathogenic strains. The two clades that are genetically isolated should be recognized as two subspecies both independently represent a continuum of possibilities range from commensal to pathogenic phenotype. Additionally, the frequent intra-clade recombinant events, often in larger fragments of over 5kb, indicates possibility of highly-efficient gene transfer mechanism depending on inheritance. Underlying molecular mechanisms that constitute such recombinant barrier between the subspecies deserve further exploration and investigations among broader microbial taxa.

IMPORTANCE

The concept of bacterial species has debated over decades. The question becomes more important today as human microbiomes and their health relevance are being studied extensively. The human microbiomes where thousands of bacterial species co-habit need to be deciphered at minute details and down to species and subspecies. In this study, we scrutinize the population genomics of *E. coli* and define two subspecies that are distinct from each other concerning physiology, ecology, and clinical features. As opposed to extensive genetic recombination within subspecies, limited genetic flux between subspecies leads to their phenotypical distinctions and separate evolution paths. We provide a key example illustrating that the divergence of a species into two subspecies depends on recombination efficiency; when the recombination efficiency becomes a barrier the species appears split into two. The *E. coli* scenario and its molecular mechanisms deserve further exploration in a broader taxa of microbes.

INTRODUCTION

Escherichia coli, best known as a ubiquitous member of the normal intestinal bacterial microflora in humans and other warm-blooded animals, is a Gram-negative bacterium of the family Enterobacteriaceae. *E. coli* persists as a harmless commensal in the mucous layer of the cecum and colon normally, whereas some variants have evolved to adapt pathogenic lifestyle that causes different disease pathologies, including pandemic and lethal episodes (1). Depending on the site of infection, pathogenic *E. coli* strains are divided into intestinal pathogenic *E. coli* (IPEC) and extraintestinal pathogenic *E. coli* (ExPEC), which are able to successfully propagate both intra- and extra-intestinally, respectively. Naturally, *E. coli* is a highly versatile species that survives in diversified ecological habitats, such as sludgy environment of lake/river banks and tidal zones, as well as human and animal intestines (2). Their environmental adaptation and lifestyle alteration, together with experimental manipulation, make the species an excellent model for studying commensalism-pathogenicity and genotype-phenotype relationships.

Phylogenetic analysis reveals that *E. coli* exhibits complex within-species sequence diversity that hinders strain classification although various typing methods including ribo-typing, MLST (multi-locus sequence typing), phylotyping, and whole genome phylogeny have been applied. Albeit complicated, the population of *E.coli* is largely acknowledged as clonal, and four major phylotypes, A, B1, B2, and D have been identified (3) which basically differ in habitat and life-style (4, 5). Phylotypes are loosely associated to phenotypical characteristics, such as antibiotic-resistance and growth rate (3), and also correlates with pathotype, as the ExPEC strains are normally part of B2 and D (6), whereas the IPEC strains belong to A, B1, and D (7). However, recent genome-wide sequencing studies have revealed dispensable/variable genes take a large part of genome plasticity contributing to biological diversity of the species. Many virulence genes, including the most lethal Shiga toxins (Stx) and carbapenem hydrolyase, are subject to frequent horizontal gene transfer (HGT), through which distinct pathogenic and resistance phenotypes are acquired (8). Therefore, extensive HGTs interrupt the connections between phenotypes and their mainstream phylogeny. Meanwhile, homologous recombination of the *E. coli* core genome is found more frequent than previously realized, which also obscures phylogeny and leads to either divergent or convergent characters (9). These observations on genetic flow challenge the clonality of the species population, and raise the concern on the emergence of

89 disastrous “superbugs” if both lethal toxins and highly-resistant genes are recombined into a new
90 strain (10, 11).

91 To understand the population structure and genetic diversity within the species *E. coli*, we utilize
92 104 high-quality complete genome sequences from diverse geographic and host range; among them,
93 45 that have animal-host information are first released from our own sequencing efforts and the rest
94 of explicitly human hosts are from public databases. Our analysis results support the view that *E. coli*
95 is predominantly clonal, and except that a few strains of intermediate minor clades, most strains
96 cluster into two major clades. Each in successful adaptation to their own ecological niche and the two
97 clades distinct in physiological features, pathogenicity, genome content, and homologue sequences;
98 we propose that they deserve a permanent distinction as two subspecies. Further HGT analysis
99 reveals that recombination is very extensive within subspecies, which homogenizes strains into
100 continuum of genome possibilities, but rather limited when it happens in cross-subspecies manner.
101 The barrier of genetic exchanges between the two subspecies maintains clonal characters of the
102 species and drives them into separate evolutionary paths.

103

104 **RESULTS AND DISCUSSION**

105 **We contribute a unique fraction of complete genome sequences for a dataset used for**
106 **phylogenetic analysis.**

107 Our dataset for an in-depth analysis is composed of 104 complete genomes which include 59
108 human-host complete genomes from public databases and 45 newly added complete genomes of
109 animal-host isolates. The latter set is selected from a world-wide collection of 202 *E. coli* strains of
110 animal hosts, which includes strains with broad diversity in geography, climate, and host range, and
111 several ECOR (*Escherichia coli* collection of reference) strains (12). Our effort includes the
112 identification of the MLST (multi-locus sequence typing) sequences for constructing a MLST–based
113 phylogeny. The MLST-based phylogeny reveals a complex partition among animal or human hosts
114 (Figure 1A), and at the end, 45 animal-host isolates is brought on to represent diverse origin and
115 genetic heterogeneity in terms of host diet, geographic distribution (Figure 1B), as well as MLST-
116 clusters for further genome sequence finishing; their full genome sequences reveal similar

chromosomal organizational features to human-host isolates, such as a uniform G+C content of 50.58%. Their genome sizes range from 4.25 to 4.94Mb and are predicted to encode 4,728 genes in average, a slightly smaller than that of pathogenic strains, but similar to commensals.

To construct the genome-based phylogeny, 7 draft genomes of strains isolated from wild environment (13), which are phylogenetically distant from host-related strains, are included as outgroups. The core genome shared by all the 111 genomes is composed of 1,095 genes and collectively 1.05 Mb in length. The maximal-likelihood phylogeny of the core genome indicates ancestor position of the environmental strains (Figure 2) and the host-related strains, regardless of human or animal origin, are mingled together. The majority of the host-related strains clusters into two major clades, each contains 56 and 31 strains respectively. The rest of 17 strains, holding closer positions to the ancestral environment strains, is split into several minor clades. Overall, this clustering pattern is largely in congruence with their phylotypes as previously reported (9, 14).

The two major clades are physiologically distinct and adaptive to each ecological niches.

The two major clades of host-related strains also exhibit distinctions in host, climate (5), and pathogenicity (Table S1). According to their distinct characters, such as movement and growth, we name them Vig (Vigorous) and Slu (Sluggish) clades just for convenience. The Vig, composed of strains of phylotype A and B1, is featured for its strains of carnivores host and tropical geographic distribution; all *E. coli* strains that have led to human pandemic infections belong to this particular clade. Although some Vig strains also survive in herbivores and cold area, and many strains are commensals, it appears that the Vig strains prefer warmer temperature, richer amino acid nutrients, and are able to propagate rapidly under ideal conditions, albeit adapt to a wide range of ecological niches. On the contrary, the Slu strains, which fall in phylotype B2, are only found in herbivores and omnivores and colder climates. Besides commensals, Slu strains occasionally cause extra-intestinal and antibiotic-resistant infections that are rarely seen among the Vig strains. These ecological features of Slu strains suggest that they are more adaptive to lower temperatures, poor amino acid nutrients, and therefore exhibit slower growth rate. This adaptation to low temperature of the Slu clade (mainly B2 strains) coincides with the report of a large population investigation (15), and facilitates transient period of epizoism and migration to extra-intestinal habitats together with their tolerance to

poor nutrition. Meanwhile, its slow growth rate increases survival rate under stresses, such as those of antibiotics, and offers better opportunities to gain antibiotics-resistant genes or functional mutations.

To confirm some of the distinct features between the clades Vig and Slu, such as optimal temperature and amino acid preference, we design a few straightforward physiological experiments with resuscitated strains in our collection (23 Vig and 15 Slu strains). First, we measure growth rates for the strains at various temperatures and find that the Vig strains grow significantly faster than the Slu strains at 37℃ and 41℃, which resemble the intestinal environment of warm-blood animals (Figure 3A). However, at lower temperature of 27℃ and 32℃, their differences are not significant. The results well explain the higher prevalence of the Vig strains and the preference of Slu strains to colder climates. Second, we compare their survival rates after heat shock at 50℃, and the faster-growing Vig strains exhibit vulnerability and poorer survival rates as compared to the Slu strains after 20 and 30 minutes of heat stress (Figure 3B), which is in congruent with the measurement of their growth rates. Third, we measure mobility of these strains in response to chemotaxis amino acid—phenylalanine—in a custom-designed microfluidic device. At each time point, more cells of the Vig strains reach the destination pool at the same chemoattraction, indicating their faster response to amino acids and higher mobility than the Slu (Figure 3C). This ability of Vig strains ensures rapid approaching toward amino acid nutrients and allocates to their carnivoral inhabitation. Finally, we calculate the strain growth rate in various carbohydrate sources relative to glucose. All *E. coli* strains grow much rapidly in monosaccharides and disaccharides than in polysaccharides. However, when compared to the Slu strains, the Vig strains grow slightly faster in monosaccharides and disaccharides, but a little slower in polysaccharides (Figure 3D). Although the difference is not significant possibly due to small sample size, it seems that the Slu strains may be more adaptive to herbivore host where polysaccharides are more abundant. From all the physiological experiments, it is apparent that the two clades have diverged from each other in phenotypical features in adaptation to distinct ecological niches as well as leading to distinct clinical features.

Genomic distinctions between the two clades correspond to their phenotypes.

The distinct physiological and ecological features between clades Vig and Slu lead to the speculation that the two clades should be distinct in genome content and homologous sequences especially in genes related to metabolism, energy, and mobility. To confirm this, we formulate a

parameter – pair-wised genome content distance – of all strains, finding that the Vig and Slu strains are clearly separated in a neighbor-joining tree that derives from the distance matrix (Figure 4A); the result indicates that the two clades have a significant number of different genes. We subsequently apply two-sample Kolmogorov-Smirnov test to all dispensable genes and identify 279 and 336 orthologues that exhibit significant enrichment in the Vig and Slu clades, respectively (Table S2). These genes represent genome content that are characteristic of the two clades; when annotated according to the COG database for their functions, as expected, over half of them are in the categories of metabolism. However, although different between the Vig and Slu clades, the distribution (presence/absence) of these characteristic genes does not vary so much with host dietary preference (Figure 4B). For metabolizing carbonates and lipids, the Vig and Slu contain a set of diversely characteristic genes. For genes of mobility related, purine and amino acid metabolic, the Vig strains are apparently richer than the Slu, which are in accordance with their adaptation to carnivore intestines (Figure 4B).

Next, we search for all virulent genes within dispensable genes. For both clades, unlike metabolic genes, the content of virulent genes varies greatly among strains. Pathogenic strains have much more virulent genes than commensals, including all kinds of toxins and iron-uptake genes. However, the two clades differ in their virulent gene content: the Vig pathogens are rich of T3SS and other secretion system, whereas the Slu pathogens have more adhesion and invasion genes that facilitates extra-intestinal infections (Figure 4C). We also scrutinize beta-lactamase genes and the notorious lethal Shiga-like toxins (Stx) that often leads to pandemic infections when transferred to *E. coli* strains (16). These genes are usually carried in plasmid, but can be inserted into chromosome by mobile elements under strong selections (17). Although both clades contain strains with narrow spectrum beta lactamases as TEM-1 and OXA-1 (18), the extended spectrum beta lactamases (ESBLs) are rarely found among Vig strains, whereas Stx exclusively shows up in Vig strains. (Figure 4D). These differences between the two clades in their genome content lead to explanation for their phenotypical and pathogenic characteristics.

In addition to gene content, we further investigate sequence variances between homologs that are shared by the two clades. First, the Vig and Slu clades are phylogenetically distinct in their core genome, and from the orthologous pairs we identify 126 and 227 non-synonymous polymorphic sites

which reside in 97 and 168 genes and are specific for the Vig and Slu clades, respectively (Table S3). These polymorphic sites exclusively found in each clade are named as lineage-associated variations or LAVs. The ratios of non-synonymous to synonymous polymorphism (dN/dS) of these LAV-containing genes are in average 0.02 and below 0.4, as if they are under strong purifying selection. We extract all metabolic genes in carbonate and amino acid pathways from the core genes and concatenate them to construct maximal-linkage tree. Clearly, genes of Vig and Slu strains cluster separately (Figure 5A), likely to contribute to their metabolic characteristics. We also investigate homologous sequences of the highly prevalent dispensable genes (orthologues present in over 80% strains of the two clades and calculate pair-wised gene distances based on amino acid sequences between orthologues. The result shows much larger distances between inter-clade pairs than intra-clade pairs (Figure 5B), indicating that each clade utilizes their own preferred alleles for highly shared dispensable genes. Finally, we check the GC content difference between the two clades, and unexpectedly find that GC% of the Vig strains are a slightly, but significantly, higher than that of the Slu strains in the core genes, dispensable genes, and whole genome (Figure 5C). We speculate that the underlying reasons are adaptation to the higher temperature habitant of the Vig clade. In any case, the difference in GC content definitely leads to a global effect on nucleic acid composition of all genes and can be an independent factor driving genome evolution (19). Apparently, all above genetic characteristics in genome content and sequence diversity between the two clades may explain in part their phenotypic/physiological differences and molecular mechanisms for adaptation to their distinct ecological niches.

Barrier to inter-clade recombination separates the clades genetically into subspecies.

E. coli is generally clonal (20, 21) and the Vig and Slu strains fit in such a framework, diverging from each other through accumulation of independent mutations and limiting exchange of genetic materials mutually. However, in previous studies, the *E. coli* recombination rate is evaluated as comparable to a mutation rate with the ratio of $r/m \approx 0.9$ (9, 22, 23). Theoretically, such a recombination rate is able to confound clonal framework and intermingles strains into an unstructured population. The question becomes how the species keeps clonal structure and sustains a relatively high recombination rate.

To scrutinize clonal status, we first evaluate general recombination rate based on the concept of homoplasy – polymorphisms shared by two or more strains but not present in their common ancestor (22, 24). Practically, homoplasic polymorphisms can only be inferred for core genome where recombination are mainly introduced through homolog sequences. Our results indicate that the *r/m* of the *E. coli* core genome is about 0.5, much lower than previously estimated (22) (Table 1). The reason is the fact that dispensable genes are more subjected to HGT and lead to over-estimated *r/m* ratio when they landed in a core genome. The large sample size of our current study results in a minimal definition of the core genome that excludes almost all the dispensable genes and thus narrows *r/m* ratio. However, the *r/m* ratios of the Vig and Slu clades are 0.949 and 0.745, much higher than the entire species (Table 1); the deviations point out the fact that there are more within-clade recombination events than that of between-clade, as it has been proposed in a previous study (9). The high rate of within-clade homologous recombination is further confirmed by an analysis based on a Bayesia-based method—BRATNextGen, which has been used for analysing homologous recombination events between and within clades on the basis of a specified degree of sequence divergence (25, 26). The result illustrates the fact that strains share more recombinant fragments within clades than cross clades (Figure 6A). Both analyses demonstrate that *E. coli* strains rarely exchange genetic material with distant relatives whereas closely related strains recombine more frequently, and thus reconciling the controversy between clonal structure and high overall recombination rate of the species.

The above estimations of recombination rate are applied for the core genome which only takes less than one-fourth of the average *E.coli* genome, however, recombination rate varies with chromosomal regions, and genes differ in their possibilities of being transferred—dispensable genome contains more mobile elements and are more prone to be horizontally transferred (27) To obtain complete understanding of genetic exchange in term of whole genome and all strains of entire species, we try to scan recombinant events across the entire genome length; recombination plays critical roles in transferring and shuffling dispensable genes that shape genome content in significant ways (28). In general, recombinant events between close-related strains are not easy to identify due to weak signals of recombined sequences (29, 30). Based on sequence alignment, we first identify all near identical fragments for each genome pair as candidate recombinant fragments (31). Since recombinant often insert genes into different location, using complete genome sequences, we also

compare synteny (linear order) of the candidate recombinant fragments and define non-syntenic fragments as results of true recombinant events. Our result shows that strains with intra-clade genome-pairs exchange more genome content (nearly 10x) than those with inter-clade pairs; the sum of total recombinant fragment lengths approaches one half of the genome in some cases and never less than 300kb which is even larger than the upper limit of inter-clade pairs (Figure 6B). The extensive intra-clade genetic exchange in the dispensable genome strengthens the clonal structure that is defined by the core genome. Furthermore, we find that intensive genetic material exchange where virulent genes can be included between commensal and pathogenic strains depends on their clade statuses: higher in within-clade events (Figure 6B).

We further correlate recombinant frequency (here we use the number of recombination fragments per genome-pair) to phylogenetic distance between paired genomes. When plotted against core genome distance (from which the core genome phylogeny has been inferred) for both Vig and Slu strains, the regression curves of recombination frequency shows a reverse-S shape, i.e., there is a rapid decline over transition between intra- and inter-clade genome-pairs (Figure 6C). Strains of the same clade recombine more frequently, which often have nearly one hundred recombinant fragments per genome, and show a decrease in the number of recombination fragments due to overlapping of larger fragments when very closely related. In fact, very close sister strains can exchange almost half of their genome. The phenomenon of chromosomal fragment transfer over 100kb or even 1Mb has been experimentally validated in other species (32, 33). However, the very efficient genetic exchange of large fragment and high frequency is not seen between clades, where such fragments rarely exceed 5kb (Figure 6D). The result implies that closely related strains (within clade) may have a unique highly-efficient molecular mechanism for recombination, whereas genetic material exchange between remote relatives (cross clade) is confined to some low-efficient ways.

Population structure of host-related strains and relationship between phylotype and phenotype

E. coli is a well-studied gram-negative bacterial species due to its importance in both clinical practice and biological research. However, biologically meaningful strain or population classification based on genotypes and phenotypes as well as other features is still a difficult task. Our study starts from phylogeny based on high-quality whole genome and, through detailed genomic analysis and

experimentation ends with the definition of two major clades – Vig and Slu. The two clades are distinct in many aspects and their clade-centric characteristics explain many ecological and clinical features. *E. coli* infections fall into two categories with different clinical outcomes and treatment strategies: 1) acute intestinal infections with various severities and 2) opportunistic infections often in extraintestinal loci and often resistant to common antibiotics. The two types infections appear to correlate with different clinical features of the two clades and their traditional phylotypes (34, 35). Genomic analysis reveals that the genetic structure underlying the physiological, ecological and clinical traits of Vig and Slu strains are a pile of characteristic genes and sequence variations, especially genes involved in metabolic pathways, mobility, and toxic or resistant phenotypes. And the separation of the two clades is caused by the limitation of between-subspecies genetic exchange or recombination, as similar scenarios found among other species of bacteria (36-38), archaea (39), and eukaryotes (40). Among the characteristic genes, the extremely virulent toxin—Shige-like toxin and the most notorious antibiotic resistant genes—ESBLs, are partitioned into the Vig and Slu with very little overlap. Although these genes are often carried by plasmids and ready to transfer among strains, it seems that strains carrying both has been reported to be rather rare (41), also supporting the between-clade recombinant barrier. Therefore Vig and Slu should be regarded as two subspecies since the recombinant barrier has genetically separated them and made them clearly divergent in all aspects of physiology, ecology and clinical significance. In clinical, identification of subspecies for a pathogenic strain will give informative clinical guidance for the treatment of the infection it caused. On the other side, the genetic boundary between commensals and pathogenic strains is not very clear. Although pathogenic strains bear much more toxic genes, a commensal strain can exchange genetic material and acquire enough virulent genes from its close pathogenic sisters, and then become pathogenic under appropriate host conditions. Intensive recombination readily alters virulent gene content and thus blurs the lines between commensal and pathogen, making them genetically undistinguishable in clinical practice (6).

Since species are “lineages evolving separately from other lineages” (42), genetic diversification and geographic separation of *E coli* subspecies, represented by the Vig and Slu clades, demonstrate an early process of microbial speciation, whose mechanisms and processes have been debated over decades (43, 44). Until recently, technological innovation, especially the invention of next-generation sequencing (NGS) technology, coupled with the emerging discipline of population genomics, has

been providing unprecedented tools and opportunities for the interrogation of molecular details on many ongoing evolutionary processes among natural microbial populations, especially those with healthcare applications. The emergence of the two *E.coli* subspecies appears initiated from distinct genetic units with functional relevance, which are incubated and frequently traded among closely related, structurally comparable, and geographically cohabitating strains through mutually beneficial mechanisms. Our recombinant analysis reveals that the within-subspecies recombination rate is much more significant than that of between-subspecies, and such a diversifying process eventually drives subspecies or their populations keeping evolving into nascent species (45). In our data, recombination has overall effects on both core and dispensable portions of the genome, and results in hundreds of characteristic genes as well as lineage-associated variations or LAVs; these genetic and functional elements form a complex background for species and its population to evolve under nature selection. Certainly, physical barriers that interrupt recombination, accelerating the process of speciation (43). In the case of the two *E coli* subspecies, both are widely spread and co-inhabiting, such as commensals in intestines of both humans and other omnivores, and our study and observations of them does not support the hypothesis of geographic isolation. Therefore, other types of physical barriers such as CRISPR-Cas system (46), restriction-modification system (47), DNA uptake signal sequences (48), and incompatible transfer mechanisms due to pili (49), which have been reported in some species, may play roles in *E. coli* sub-speciation or speciation but remain to be elucidated. Our results highlight rate difference between intra-subspecies and inter-subspecies recombination as a barrier possibly due to less functional benefits. Some high-efficiency mechanisms, such as distributive conjugal transfer, has been reported in a species of Mycobacteria, which is able to transfer fragments over 100kb at one time but lose function when such genetic exchange happened between remote relatives (50, 51). On the other hand, low-efficiency mechanisms, such as phage (52) or transposon (53), usually transfer short fragments at lower frequency (54), but may still work when happening across subspecies due to broader host range. These mechanisms underlying cross-subspecies recombinant barrier deserve further exploration and should be investigated in a broad range of microbial taxa for better understanding of microbial population structural dynamics and speciation.

CONCLUSION

Based on an unprecedented dataset, we thoroughly studied the population structure and dynamics of *E. coli*, including genetic diversity, habitant divergence, physiological features, recombination rate, and gene flow. We defined two *E. coli* subpopuations among large number of isolates and suggest that they appear to be distinct subspecies that have evolved to bear characteristic gene content and sequence variance, which lead to distinct physiological, ecological, and clinical characteristics. There is an apparent barrier of recombinant between the two subspecies, which drive their genetic diversification. Although the underlying mechanism still needs further demonstration, novel molecular mechanisms differentiating intra- and inter-subspecies genetic material exchange may exist. The discovery of such mechanisms and confirmation in broad range of microbial taxa will surely deepen our understanding of the process of bacterial speciation.

MATERIAL AND METHODS

Strains and MLST typing. A world-wide collection of 202 *Escherichia coli* strains from vertebrate hosts (12) was kindly provided by Professor Shulin Liu (Genomics Research Center, Harbin Medical University, Harbin, China; Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, Canada). Genomic DNA were extracted using a Qiagen DNeasy kit (Qiagen) for 172 successfully resuscitated isolates. PCR and Sanger sequencing were performed for 16S rDNA, MLST genes (seven housekeeping genes of adk, fumC, gyrB, icd, mdh, purA, and recA, see details at http://mlst.ucc.ie/) for these strains. Then five strains whose 16S rDNA sequence showed <97% identity with *E. coli* reference or hit best to other species and 18 strains whose MLST alleles were failed to be identified by this method were removed, leaving 149 animal-host strains for further analysis. We further included 59 complete genomes of human-host deposited in NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) as of July 2013 and seven draft genomes published by a study of environmental strains (13) from ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT/, then *in silico* identified their MLST loci with BLAST. Together, we aligned the seven MLST fragments of all 213 strains with ClustalW and concatenated them for inference of maximum-likelihood phylogeny using RaxML with model GTR and 1,000 bootstraps.

Genome sequencing, annotation and phylogeny inference. Genomic DNA libraries of 45 selected representative strains were prepared using a NEBNext DNA Library Prep Kit and sequenced on a Hiseq 2000 for 2 × 100 bp run at the Beijing Institute of Genomics (Beijing, P.R.China). The raw reads were quality filtered and trimmed using SolexaQA (-p 0.01), with an average coverage of 150× and assembled with SOAPdenovo (S) (55), and then the assemblies were scaffolded into circular genomes with GAAP (http://gaap.big.ac.cn), which is based on synteny of core genes on genomic scale, and assist assembly of high quality genomes (27). For these genomes, protein-coding genes were predicted by using GeneMark.hmm with a pre-trained model and annotated using BLAST against the COG database. Gene sequences were mapped to metabolic pathways by using BLAST against KEGG GENES for KEGG Orthology assignment using the KEGG automatic annotation server. The identification of orthologous genes was performed with pan-genome analysis pipeline (PGAP) (56) with identity and coverage of pairwise genes set at 0.7. Orthologues common to all 111 strains (including 45 animal-host, 59 human-host and seven environment strains) are identified as core genes, and others (orthologues shared by a portion of strains) as dispensable genes. The 1,095 single-copy core genes were concatenated into core genome and used to construct phylogenetic tree. We aligned protein sequences of each core genes using ClustalW, and traced them back to nucleic acid sequences. The maximum-likelihood phylogeny was inferred on the basis of SNPs in the core genome based on RaxML in the model GTR+G with 1,000 bootstraps. Phylogenetic trees were viewed and modified by using FigTree and EvolView (57). Phylotypes of all strains were identified *in silicon* according to the presence of three phylotype-specific genes or fragments as previously described (58).

Measurement of growth rate, survival rate, and mobility. Strains were resuscitated and cultured in pre-filtered LB broth. Unless otherwise specified, each strain was normalized by cultivation at 17°C in LB overnight with shaking at 200 rpm, and OD600nm was measured for each culture in 96-well plate (triplicates for each strain with a blank control) on an Infinite® 200 PRO Microplate Reader (Tecan) every half hour.

Growth rate in diverse temperature and carbonate source. Overnight cultures were 1:50 diluted with LB broth and incubated at 27, 32, 37, and 41°C or change medium supplemented with glucose, starch,

glycogen, heparin, cellulose, fructose, lactose, and maltose as sole carbon source. OD_{600} was monitored until the cultures reach 0.6.

Survival rate at heat stress. Overnight cultures of each testing strain were adjusted to OD_{600}=0.1. Immediately, and 50µl of inoculum was diluted 1:4 with LB broth and incubated at 50°C for 0min, 10 min, 20min, and 30min. OD_{600} was monitored until the cultures reach 0.1. The number of living cells in each initial aliquot relative to a negative control was calculated as $e^{-\Delta t}$, where Δt is the time to OD_{600}=0.1. The survival rate was defined as the ratio of live cell number after heat shock to that of negative control.

Mobility and response to chemoattractive amino acids. Each strain was evaluated on a custom-designed PDMS microfluidic device as previously described (59). Briefly, there were two pools—sink and source pool—which were connected with a 600µm channel. At the end of the channel adjacent to the source pool, a thin layer of precast gel obstructed cells into the source pool, in front of which a 400µm2 observation channel was set there for cell counting. Normalized cells of single clones were re-suspended in an amino acid-free basal medium with OD_{600}=0.1. Immediately, 30 µl of cells and 30 µl of chemotaxis solution (with 100 µM phenylalanine in basal medium) were injected in the sink and source pool respectively. Cells reached the observation channel was immediately observed and auto-tracked by using a Nikon Ti-E inverted microscope system every 5 min for 1 h. The cell count of observation channel in each images, which serves as an indicator of cell mobility in response to chemoattractive amino acid, was automatically measured by using ImageJ software.

Genome content similarity, gene distribution, and homologue distance between Vig and Slu. The similarities between paired genomes were calculated on the basis of the Bray–Curtis dissimilarity index. A dissimilarity index of d is calculated as 1 – [2 * Sij/(Si + Sj)], where Sij is the number of dispensable genes shared by strains i and j, and Si and Sj are the numbers of dispensable genes in strains i and j, respectively. Pairwise dissimilarity indices were used to construct a distance matrix, which was used to construct the neighbour-joining tree, and genome similarity (1 − d) was used to produce a heatmap. Then we applied two-sample Kolmogorov-Smirnov test for each dispensable orthologue on its balanced distribution/presence in Vig and Slu. When p <0.01, the orthologue was significantly enriched in Vig or Slu. Antibiotic-resistance genes were identified by using BLAST against CBMAR database (http://14.139.227.92/mkumar/lactamasedb) with E value of $10^{-5}$and best hit.

Similarly, Shiga-like toxin genes were identified by BLAST against the protein sequences of that from *E. coli* O157:H7 (lcl|AB015056.1_prot_BAA88123.1_1 for A-subunit and lcl|AB015056.1_prot_BAA88124.1_2 for B-subunit) downloaded from NCBI database. To calculate pairwise sequence distances for protein between homologs, individual orthologues of high-prevalence dispensable genes (present in >80% of strains) were first aligned with ClustalW, and then calculated pairwise distance of genes in one orthologue from various strains of one species by using Protdist in the Phylip package.

Recombination inference. We calculated *r/m* statistics for the core genome of *E. coli* using a computational method PHI which recognize homoplasic sites of amino acid sequences in a 1-kb overlapping window. We used all the default parameters and collected homoplasic sites in windows of $p<0.01$ as polymorphism sites caused by recombination (*r*) and the others as those caused by mutation (*m*). We also inferred recombination fragments of core genome with a Bayesian algorithm-based method—BratNextGen (25). The strain clusters were a priori defined on the basis of a PSA tree. In the default procedure, alpha was set at 3.58. One hundred iterations were performed until the model parameters converged, and the significances of inferred recombinant fragments ($p< 0.05$) were assessed using 100 replicate permutations of sites in the genome.

We utilized a programme— gmos (Genome MOsaic Structure) to compute local alignments between paired query and subject genomes and reconstructed the query mosaic structure of recombinant fragments over 600bp (31). These fragments, although almost identical caused by very recent recombination, still held the possibility of vertically inherited under strong selection, and thus only regarded as candidate recombinant fragments. We also identified fragments that changed their genome locations as parsimony recombination fragments for each genome pair. To achieve this, we assigned consecutive number for each candidate fragment along their order in subject genome. We also recorded their relative order in the query genome as a new sequence where we identified the longest increasing/decreasing subsequence as fragments that keep their original locations, and the other fragments were recognized as parsimonious recombination fragments.

Data Availability. The final 45 genome sequences we contributed were deposited in GenBank under the accession numbers CP012758~CP012800, CP012806, and CP012807.

LIST OF ABBREVIATIONS

dN/dS, ratios of non-synonymous to synonymous polymorphism

ECOR, *Escherichia coli* collection of reference

ESBLs, extended spectrum beta lactamases

ExPEC, extraintestinal pathogenic *E. coli*

HGT, horizontal gene transfer

IPEC, intestinal pathogenic *E. coli*

LAVs, Linage Associated Variations

MLST, multi-locus sequence typing

NGS, next-generation sequencing

r/m, ratio of polymorphisms caused by recombination to mutation

Slu Sluggish clade

Stx, Shiga toxins

Vig Vigorous clade

DECLARATION

Availability of data and materials: The final 45 complete genomes contributed by our laboratory has been deposit in the GenBank under the accession numbers CP012758~CP012800, CP012806, and CP012807.

Competing interests: The authors declare no competing interests.

Funding: This work was supported by the National Scientific Foundation of China [31470180, 31471237, and 30971610].

Author contribution: J.Y. and Y.K. conceived the project and led the writing; S.L, provided the strains; L.Y., Z.H., X.J., R.F., Y.Y., and C.L. compiled the data; and C.F., Z.G., X.J., X.G., Q.M., Y.Z., J.W., J.X., and S.H. analysed the data. All authors contributed to the writing and/or intellectual development of the manuscript.

Acknowledgement: Acknowledgement to W. Chen who provided manuscript feedback.

REFERENCES

1. Leimbach A, Hacker J, Dobrindt U. 2013. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Curr Top Microbiol Immunol 358:3-32.
2. Anonymous. 2017. E.coli (Escherichia coli) | E.coli | CDC, *on* Centers for Disease Control and Prevention. https://www.cdc.gov/ecoli/. Accessed
3. Gordon DM. 2004. The Influence of Ecological Factors on the Distribution and the Genetic Structure of Escherichia coli. EcoSal Plus 1.
4. Carlos C, Pires MM, Stoppe NC, Hachich EM, Sato MI, Gomes TA, Amaral LA, Ottoboni LM. 2010. Escherichia coli phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. BMC Microbiol 10:161.
5. Gordon DM, Cowling A. 2003. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. Microbiology 149:3575-86.
6. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E. 1999. The link between phylogeny and virulence in Escherichia coli extraintestinal infection. Infect Immun 67:546-53.
7. Pupo GM, Karaolis DK, Lan R, Reeves PR. 1997. Evolutionary relationships among pathogenic and nonpathogenic Escherichia coli strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect Immun 65:2685-92.
8. Martinez-Castillo A, Muniesa M. 2014. Implications of free Shiga toxin-converting bacteriophages occurring outside bacteria for the evolution and the detection of Shiga toxin-producing Escherichia coli. Front Cell Infect Microbiol 4:46.
9. Didelot X, Meric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics 13:256.
10. Zhang L, Levy K, Trueba G, Cevallos W, Trostle J, Foxman B, Marrs CF, Eisenberg JN. 2015. Effects of selection pressure and genetic association on the relationship between antibiotic resistance and virulence in Escherichia coli. Antimicrob Agents Chemother 59:6733-40.
11. Pitout JD. 2012. Extraintestinal Pathogenic Escherichia coli: A Combination of Virulence with Antibiotic Resistance. Front Microbiol 3:9.
12. Souza V, Rocha M, Valera A, Eguiarte LE. 1999. Genetic structure of natural populations of Escherichia coli in wild hosts on different continents. Appl Environ Microbiol 65:3373-85.
13. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A 108:7200-5.
14. Kaas RS, Friis C, Ussery DW, Aarestrup FM. 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics 13:577.

15. Escobar-Paramo P, Grenet K, Le Menac'h A, Rode L, Salgado E, Amorin C, Gouriou S, Picard B, Rahimy MC, Andremont A, Denamur E, Ruimy R. 2004. Large-scale population structure of human commensal Escherichia coli isolates. Appl Environ Microbiol 70:5698-700.
16. Tarr PI, Gordon CA, Chandler WL. 2005. Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. Lancet 365:1073-86.
17. Beyrouthy R, Robin F, Delmas J, Gibold L, Dalmasso G, Dabboussi F, Hamze M, Bonnet R. 2014. IS1R-mediated plasticity of IncL/M plasmids leads to the insertion of bla OXA-48 into the Escherichia coli Chromosome. Antimicrob Agents Chemother 58:3785-90.
18. Poirel L, Naas T, Nordmann P. 2010. Diversity, epidemiology, and genetics of class D beta-lactamases. Antimicrob Agents Chemother 54:24-38.
19. Wu H, Fang Y, Yu J, Zhang Z. 2014. The quest for a unified view of bacterial land colonization. ISME J doi:10.1038/ismej.2013.247.
20. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8:207-17.
21. Dobrindt U. 2005. (Patho-)Genomics of Escherichia coli. Int J Med Microbiol 295:357-71.
22. Bobay LM, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. Proc Natl Acad Sci U S A 112:8893-900.
23. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S, Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha EP, Denamur E. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5:e1000344.
24. Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, Charlesworth J, Golubchik T, Ip CL, Godwin H, Fung R, Peto TE, Walker AS, Crook DW, Wilson DJ. 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat Commun 5:3956.
25. Castillo-Ramirez S, Corander J, Marttinen P, Aldeljawi M, Hanage WP, Westh H, Boye K, Gulay Z, Bentley SD, Parkhill J, Holden MT, Feil EJ. 2012. Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. Genome Biol 13:R126.
26. McNally A, Cheng L, Harris SR, Corander J. 2013. The evolutionary path to extraintestinal pathogenic, drug-resistant Escherichia coli is marked by drastic reduction in detectable recombination within the core genome. Genome Biol Evol 5:699-710.
27. Kang Y, Gu C, Yuan L, Wang Y, Zhu Y, Li X, Luo Q, Xiao J, Jiang D, Qian M, Ahmed Khan A, Chen F, Zhang Z, Yu J. 2014. Flexibility and symmetry of prokaryotic genome rearrangement reveal lineage-associated core-gene-defined genome organizational frameworks. MBio 5:e01867.
28. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472-82.
29. Ravenhall M, Skunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. PLoS Comput Biol 11:e1004095.
30. Nielsen KM, Bohn T, Townsend JP. 2014. Detecting rare gene transfer events in bacterial populations. Front Microbiol 4:415.
31. Domazet-Loso M, Domazet-Loso T. 2016. gmos: Rapid Detection of Genome Mosaicism over Short Evolutionary Distances. PLoS One 11:e0166602.
32. Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN. 2014. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. MBio 5:e01355-14.

33. Boritsch EC, Khanna V, Pawlik A, Honore N, Navas VH, Ma L, Bouchier C, Seemann T, Supply P, Stinear TP, Brosch R. 2016. Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. Proc Natl Acad Sci U S A 113:9876-81.
34. Bukh AS, Schonheyder HC, Emmersen JM, Sogaard M, Bastholm S, Roslev P. 2009. Escherichia coli phylogenetic groups are associated with site of infection and level of antibiotic resistance in community-acquired bacteraemia: a 10 year population-based study in Denmark. J Antimicrob Chemother 64:163-8.
35. Vading M, Kabir MH, Kalin M, Iversen A, Wiklund S, Naucler P, Giske CG. 2016. Frequent acquisition of low-virulence strains of ESBL-producing Escherichia coli in travellers. J Antimicrob Chemother 71:3548-3555.
36. Huang CL, Pu PH, Huang HJ, Sung HM, Liaw HJ, Chen YM, Chen CM, Huang MB, Osada N, Gojobori T, Pai TW, Chen YT, Hwang CC, Chiang TY. 2015. Ecological genomics in Xanthomonas: the nature of genetic adaptation with homologous recombination and host shifts. BMC Genomics 16:188.
37. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science 336:48-51.
38. Zwick ME, Joseph SJ, Didelot X, Chen PE, Bishop-Lilly KA, Stewart AC, Willner K, Nolan N, Lentz S, Thomason MK, Sozhamannan S, Mateczun AJ, Du L, Read TD. 2012. Genomic characterization of the Bacillus cereus sensu lato species: backdrop to the evolution of Bacillus anthracis. Genome Res 22:1512-24.
39. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. 2012. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol 10:e1001265.
40. Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci U S A 108:2831-6.
41. Day M, Doumith M, Jenkins C, Dallman TJ, Hopkins KL, Elson R, Godbole G, Woodford N. 2017. Antimicrobial resistance in Shiga toxin-producing Escherichia coli serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015. J Antimicrob Chemother 72:145-152.
42. De Queiroz K. 2007. Species concepts and species delimitation. Syst Biol 56:879-86.
43. Krause DJ, Whitaker RJ. 2015. Inferring Speciation Processes from Patterns of Natural Variation in Microbial Genomes. Syst Biol 64:926-35.
44. Doolittle WF, Papke RT. 2006. Genomics and the bacterial species problem. Genome Biol 7:116.
45. Nosil P, Feder JL. 2012. Genomic divergence during speciation: causes and consequences. Philos Trans R Soc Lond B Biol Sci 367:332-42.
46. Samson JE, Magadan AH, Moineau S. 2015. The CRISPR-Cas Immune System and Genetic Transfers: Reaching an Equilibrium. Microbiol Spectr 3:Plas-0034-2014.
47. Pleska M, Qian L, Okura R, Bergmiller T, Wakamoto Y, Kussell E, Guet CC. 2016. Bacterial Autoimmunity Due to a Restriction-Modification System. Curr Biol 26:404-9.
48. Frye SA, Nilsen M, Tonjum T, Ambur OH. 2013. Dialects of the DNA uptake sequence in Neisseriaceae. PLoS Genet 9:e1003458.
49. Cehovin A, Simpson PJ, McDowell MA, Brown DR, Noschese R, Pallett M, Brady J, Baldwin GS, Lea SM, Matthews SJ, Pelicic V. 2013. Specific DNA recognition mediated by a type IV pilin. Proc Natl Acad Sci U S A 110:3065-70.
50. Gray TA, Krywy JA, Harold J, Palumbo MJ, Derbyshire KM. 2013. Distributive conjugal transfer in mycobacteria generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus. PLoS Biol 11:e1001602.
51. Wang J, Parsons LM, Derbyshire KM. 2003. Unconventional conjugal DNA transfer in mycobacteria. Nat Genet 34:80-4.

627 52. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. 2011. Genomics of bacterial and
628 archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev 75:610-
629 35.
630 53. Griffiths AJF MJ, Suzuki DT, et al. 2000. Prokaryotic transposons, An Introduction to Genetic
631 Analysis 7th edition. :W. H. Freeman, New York.
632 54. Mortimer TD, Pepperell CS. 2014. Genomic signatures of distributive conjugal transfer
633 among mycobacteria. Genome Biol Evol 6:2489-500.
634 55. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast
635 tool for short read alignment. Bioinformatics 25:1966-7.
636 56. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. 2012. PGAP: pan-genomes analysis pipeline.
637 Bioinformatics 28:416-8.
638 57. Zhang H, Gao S, Lercher MJ, Hu S, Chen WH. 2012. EvolView, an online tool for visualizing,
639 annotating and managing phylogenetic trees. Nucleic Acids Res 40:W569-72.
640 58. Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the Escherichia
641 coli phylogenetic group. Appl Environ Microbiol 66:4555-8.
642 59. Si G, Yang W, Bi S, Luo C, Ouyang Q. 2012. A parallel diffusion-based microfluidic device for
643 bacterial chemotaxis analysis. Lab Chip 12:1389-94.

644

645 **TABLES AND FIGURES LEGENDS**

646 **Table 1. Recombination rate inferred based on homoplasy**

	No. str.	Core genome (Mb)	No. of genes	Total polymorphism sites	Homoplasy sites	*r/m*	Source
E. coli	111	1.05	1,095	12,185	4,076	0.503	This study
Vig	56	1.05	1,095	3,293	1,603	0.949	This study
Slu	31	1.05	1,095	3,578	1,528	0.745	This study
E. coli	19	3.20	3,022	242,688	116,288	0.92	(22)

647

648 **Figure 1. Phylogenetic and geographic diversity of strains used for this study.**

649 (A) Maximal-linkage tree based on concatenated sequences of seven MLST loci of strains from
650 animal/human host and environment. Stains labelled with blue circle are selected for genome
651 sequencing. (B) Geographic and host distributions of all host-related strains.

652 **Figure 2. Maximum-likelihood phylogeny of *E. coli* strains.**

653 The ML phylogeny based on core genome polymorphisms which structures host-related strains into
654 two major clades and a few minor clades. The constitution of the two major clades is generally
655 congruent with traditional phylotype classification with a few exceptions. Each *E. coli* strain is labelled

as phylotype (coloured branch), origin (coloured inner stripe), pathotype (coloured outer stripe) and clade (Slu, the sluggish clade; Vig, the vigorous clade).

Figure 3. Physiological features of the Vig and Slu clades.

Physiological experiments for the Vig and Slu strains show their distinct features that explain their ecology distinctions. Asterisks indicate significant difference (*, $p < 0.05$; **, $p < 0.01$) between clades on the basis of *t*-test. (A) A growth rate test under variable temperatures around the normal range. (B) Survival rate test after heat stress (50℃, 20 and 30 min). (C) Mobility test in response to chemo-attraction of amino acids. (D) Relative growth rate in medium of various carbonates to glucose.

Figure 4. Genome content of strains in Vig and Slu.

(A) Genome similarity among strains. The heatmap shows pairwise genome similarity of all strains. Colours indicate scale of genome similarity, as shown in the left legend. Left to the heatmap shows the neighbour-joining tree of the strains according to their genome similarities. The Vig and Slu strains are indicated with coloured branches. (B) Average number of characteristic genes in each category of metabolism with diverse host diet, which are generally determined by clade affiliation rather than host diet. (C) Average number of virulent genes in commensal or pathogenic strains, which is influenced by both clade affiliation and pathogenicity. (D) Average number of antibiotic-resistant and Shiga-like toxin genes shows that the two types of virulent genes are clearly separated into different clades.

Figure 5. Sequence variation in shared genes of the Vig and Slu clades.

(A) Maximal-likelihood tree of concatenated genes in metabolism of amino acid (left) and carbohydrates (right) shared by all Vig and Slu strains. (B) Distribution of paired protein sequence distance between homologs for each orthologue. The red, blue, and purple lines indicate Vig within-clade pairs (Vig–Vig), Slu within-clade pairs (Slu–Slu), and between-clade pairs (Vig–Slu), respectively. (C) GC content of whole genome (left), core genes (central), and dispensable genes (right) of the two clades. Asterisks indicate significant difference (*, $p < 0.05$; **, $p < 0.01$) between clades on the basis of *t*-test.

Figure 6. Recombination in the Vig and Slu clades.

682 (A) Homologous recombination in the core genome of 111 *E. coli* strains inferred based on
683 BratNextGen. Seven clusters are a priori that is defined according to the PSA topology. Color bars
684 indicate recombined regions with various PSA cluster origins. Thick colored regions indicate
685 recombination hotspots. Gray bars indicate gaps in alignment. (B) Distribution of total length of
686 recombination between pairs of strains. The red, blue, and purple lines indicate Vig within-clade pairs
687 (Vig–Vig), Slu within-clade pairs (Slu–Slu), and between-clade pairs (Vig–Slu), respectively. Insets
688 are distributions of recombination length between commensal/pathogenic pairs (dotted line) compared
689 with intra-clade pairs (solid line) for Vig (upper) and Slu (lower). (C) Recombination frequency against
690 genome distance for Vig (left) and Slu (right) strains. Red, blue, and green dots stand for pairing with
691 Vig, Slu, minor clades, respectively, and the purple dots are Vig-Slu pairs. (D) Distributions of
692 recombinant fragment length. Recombinant fragments when cross-clade never exceed 5kb, when
693 within-clade recombinant fragments exhibiting right-shift peak overriding the 5kb-limit. The red, blue,
694 and purple lines indicate fragments of within-clade and inter-clade fragments, respectively.

695

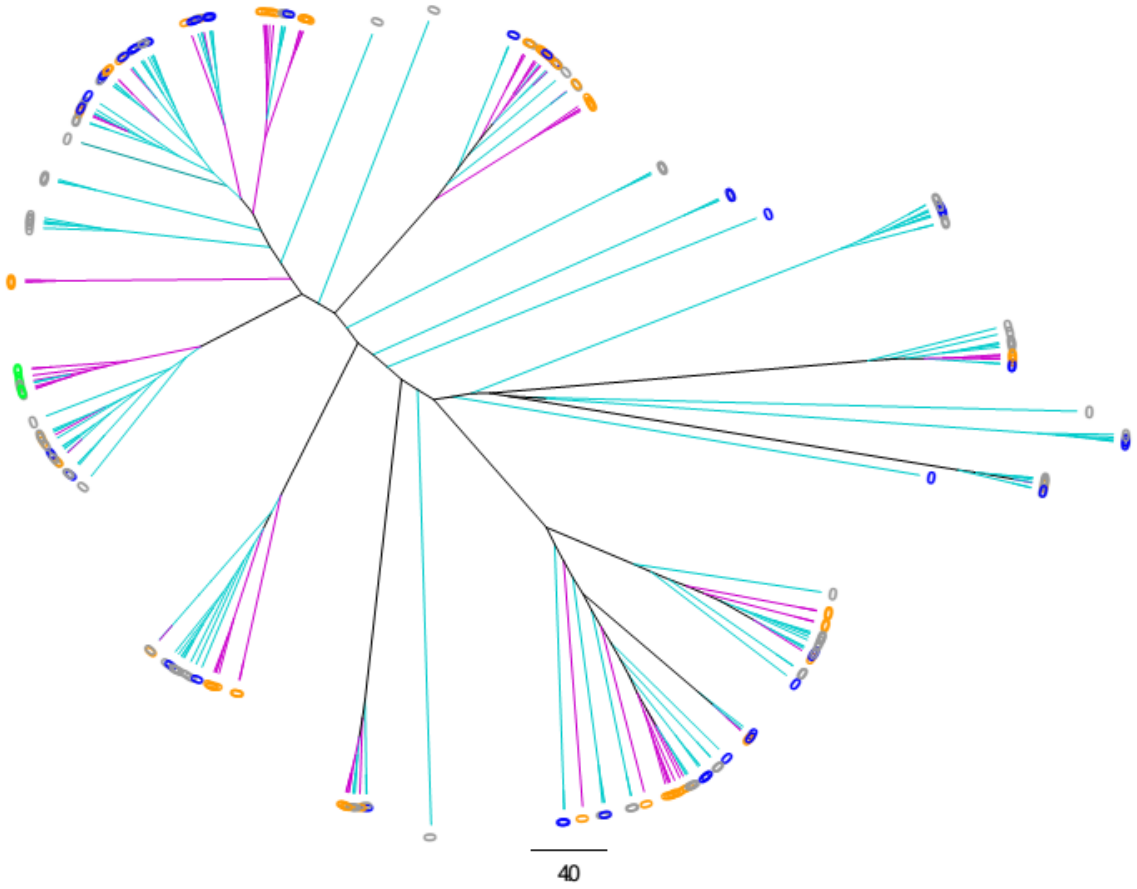
696 **ADDITIONAL FILES**

697 **Table S1. *E.coli* strains used in this study**

698 **Table S2. Characteristic genes enriched in clade Vig and Slu**

699 **Table S3. List of LAV-containing genes of clade Vig and Slu**

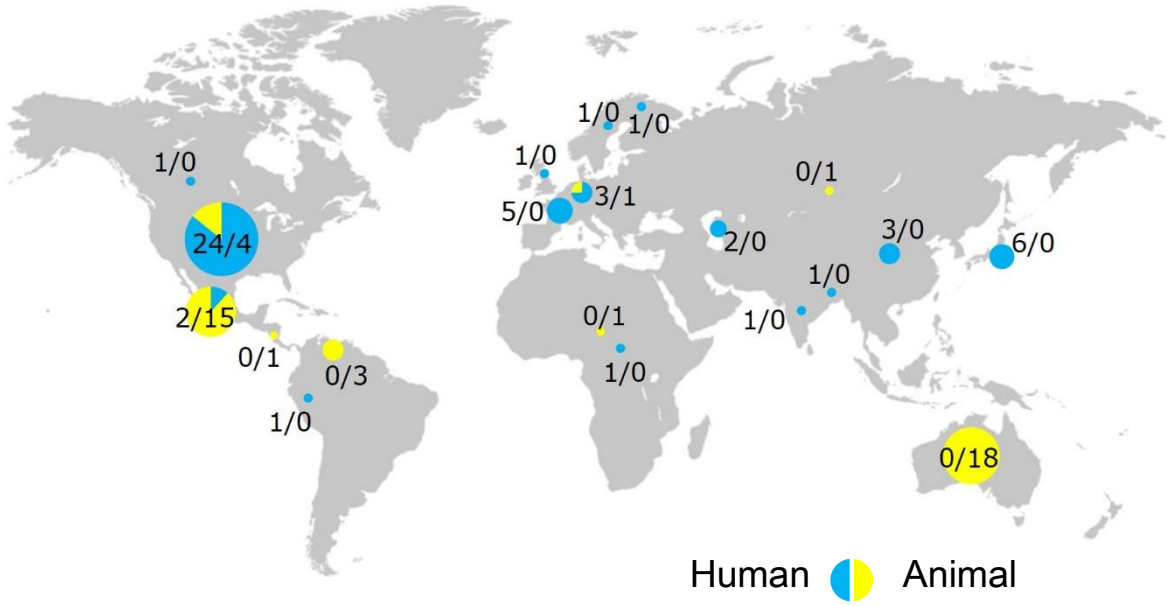
A

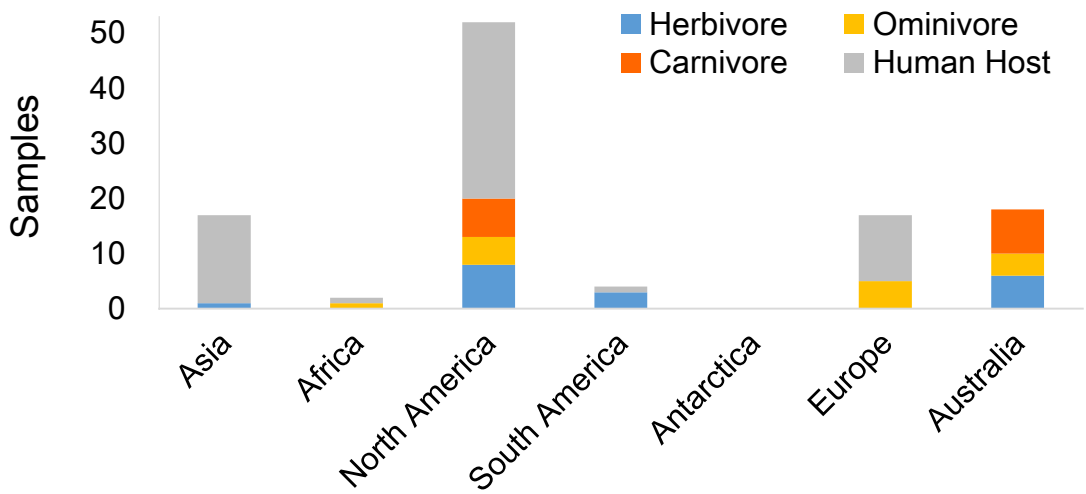


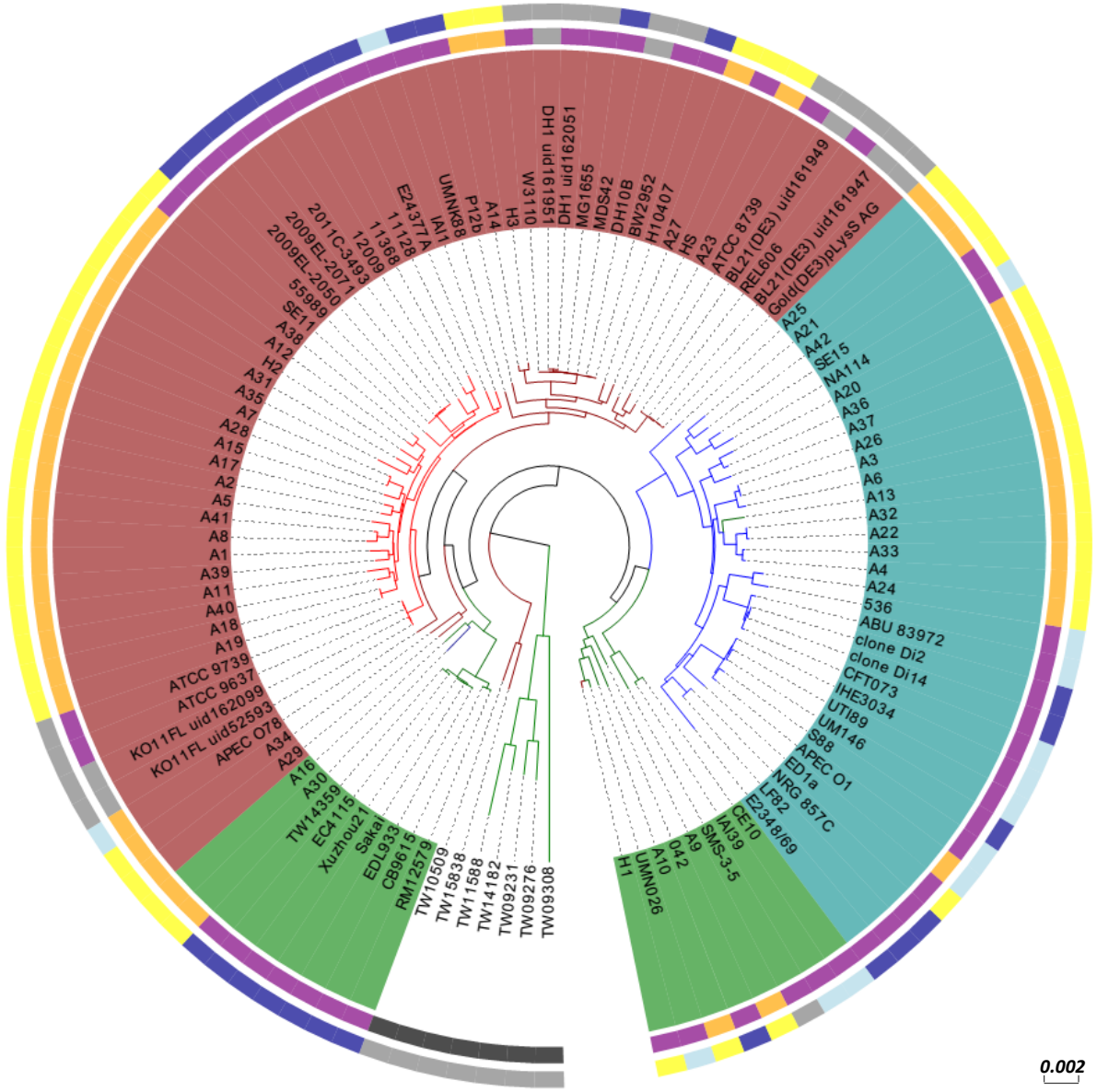
Samples

- Our collection (149 str.)
  - Self-sequenced (45 str.)
- Public collection (66 str.)
  - NCBI deposited (59 str.)
  - Drafts of environmental strains (7 str.)

B







Clade(Leaf)

- Vig
- Slu
- Minor Clades

Phylotype (Branch)

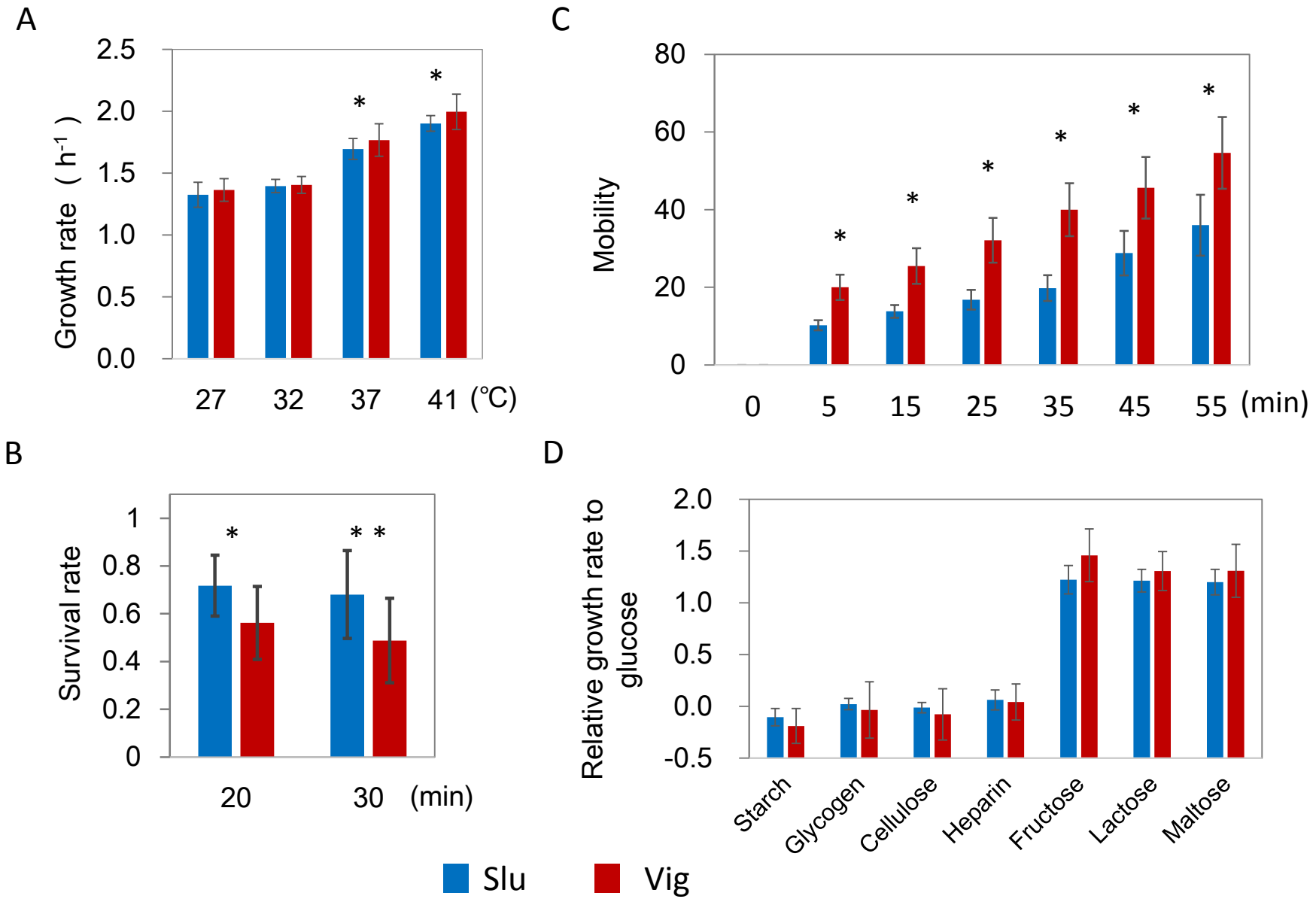
- A
- B1
- B2
- D

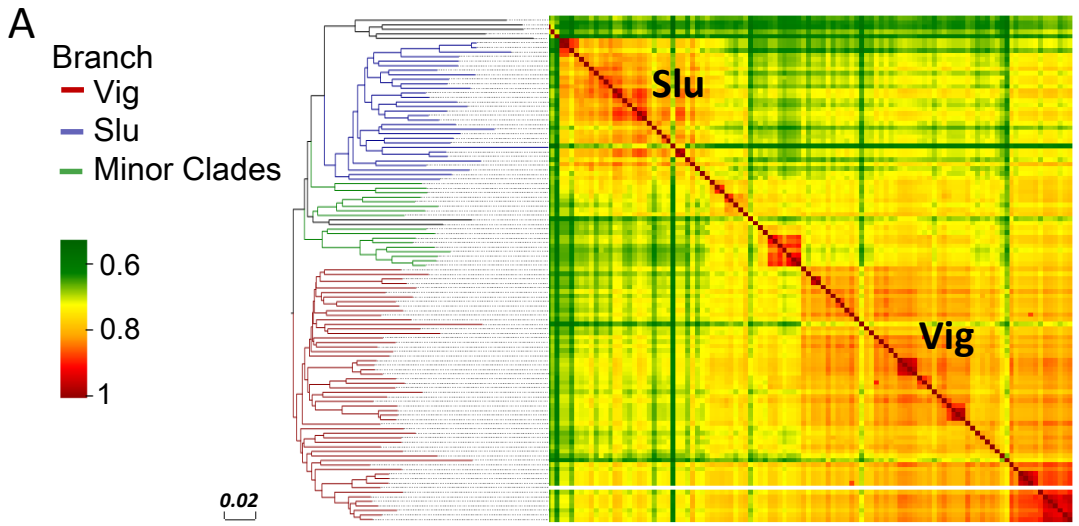
Source/Host (Inner Strip)

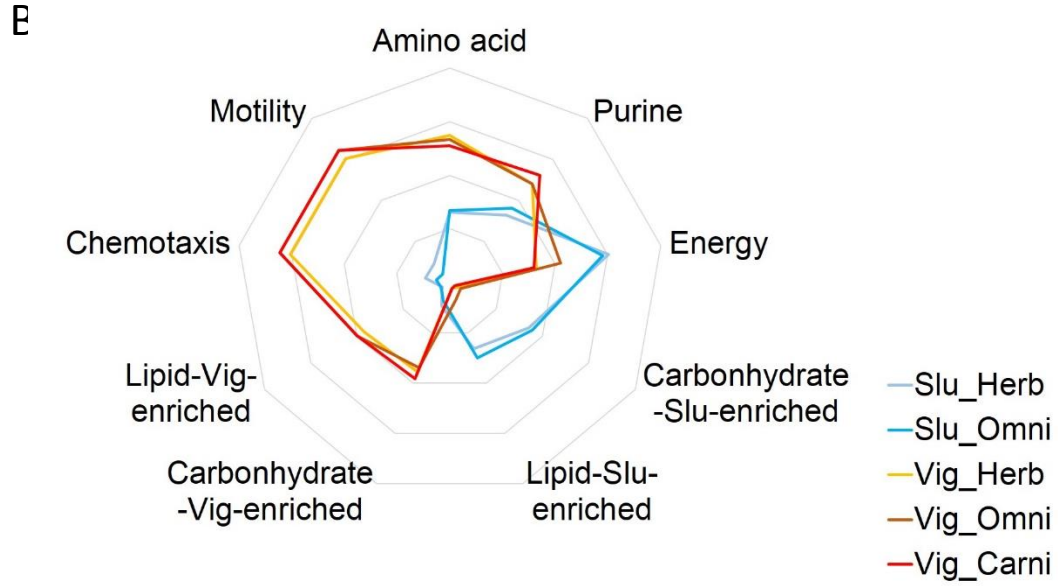
- Human
- Animal
- Environment
- Engineered strains

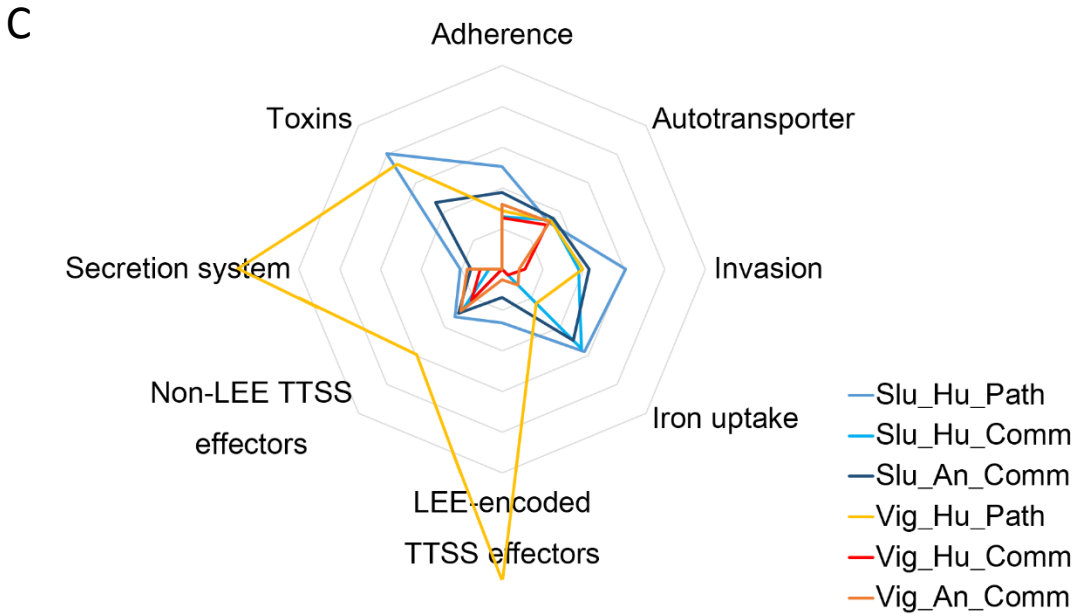
Pathogenicity (Outer Strip)

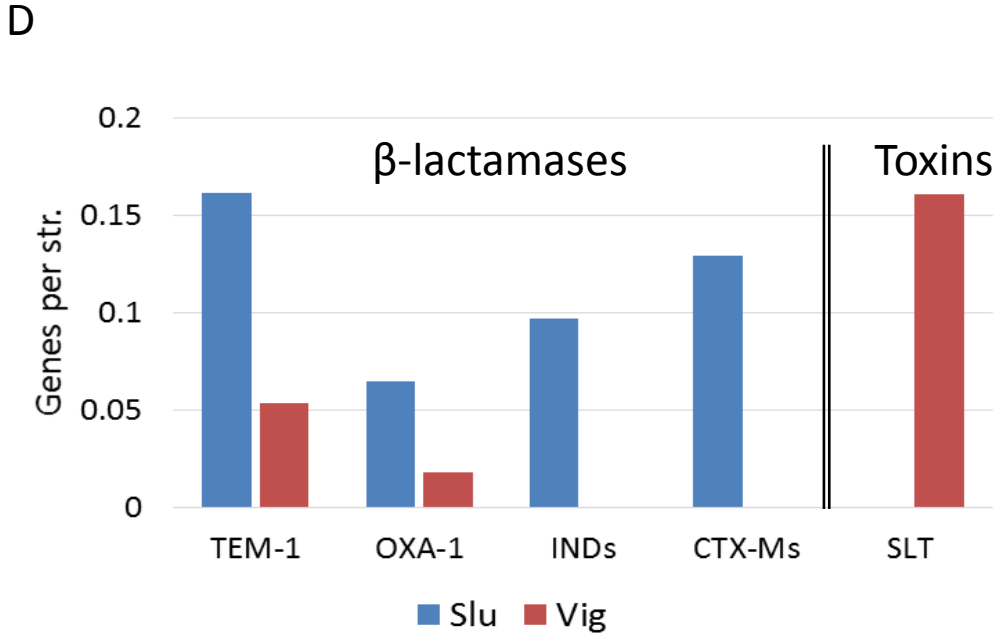
- Commensal
- IPEC
- ExPEC
- N/A


A
Growth rate (h-1)
2.5
2.0
1.5
1.0
0.5
0.0
27
32
37
41 (℃)
C
Mobility
80
60
40
20
0
0
5
15
25
35
45
55 (min)
B
Survival rate
1
0.8
0.6
0.4
0.2
0
20
30 (min)
D
Relative growth rate to glucose
2.0
1.5
1.0
0.5
0.0
-0.5
Starch
Glycogen
Cellulose
Heparin
Fructose
Lactose
Maltose
Slu
Vig

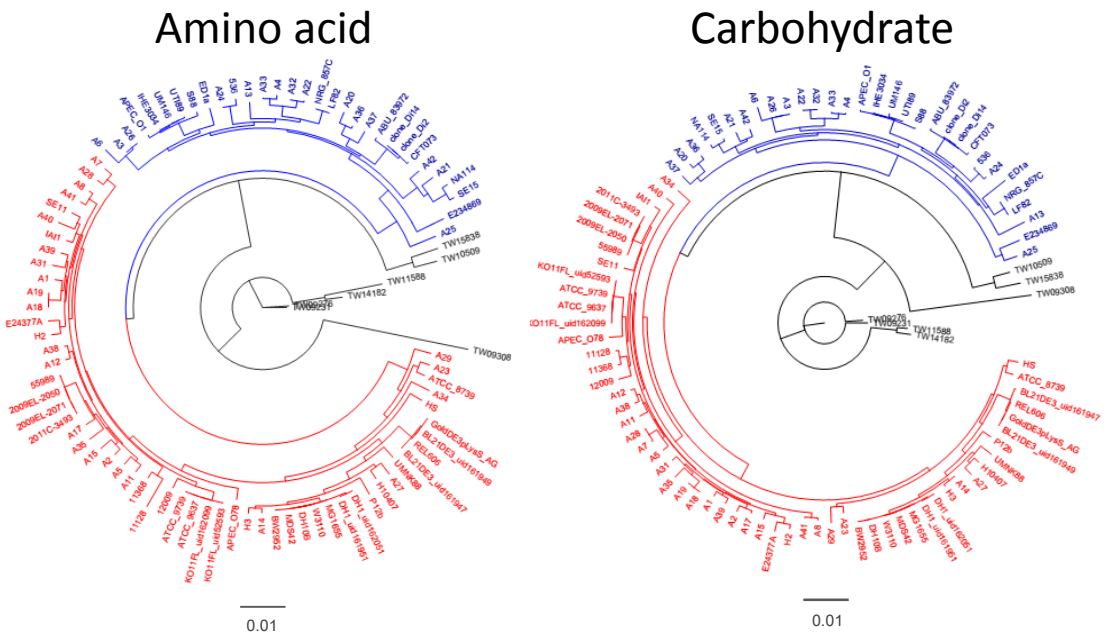

A
Branch
Vig
Slu
Minor Clades
0.6
0.8
1
Slu
Vig
0.02


B
Amino acid
Purine
Energy
Carbonhydrate
-Slu-enriched
Lipid-Slu-
enriched
Carbonhydrate
-Vig-enriched
Lipid-Vig-
enriched
Chemotaxis
Motility
Slu_Herb
Slu_Omni
Vig_Herb
Vig_Omni
Vig_Carni

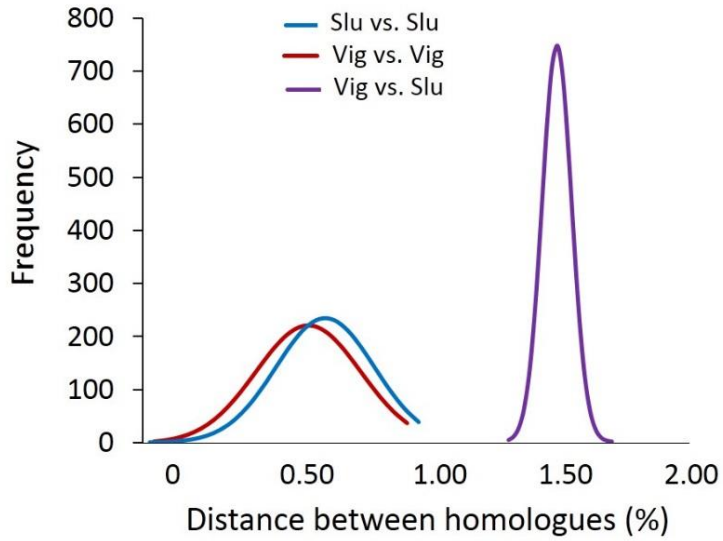

C
Adherence
Autotransporter
Invasion
Iron uptake
LEE-encoded
TTSS effectors
Non-LEE TTSS
effectors
Secretion system
Toxins
Slu_Hu_Path
Slu_Hu_Comm
Slu_An_Comm
Vig_Hu_Path
Vig_Hu_Comm
Vig_An_Comm


D
β-lactamases
Toxins
Genes per str.
0.2
0.15
0.1
0.05
0
TEM-1
OXA-1
INDs
CTX-Ms
SLT
Slu
Vig

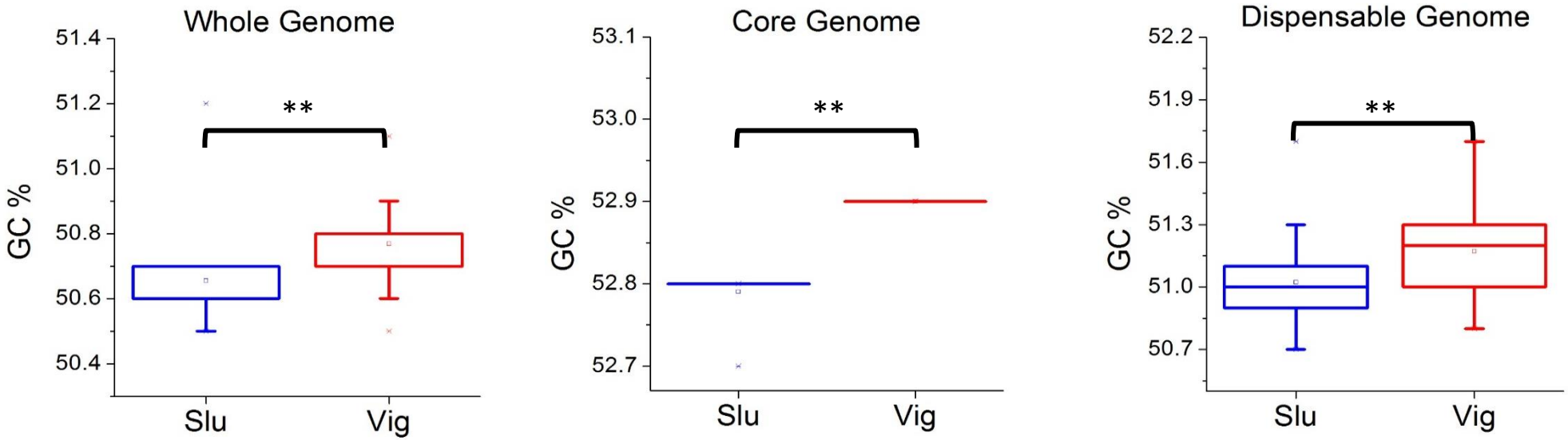
A

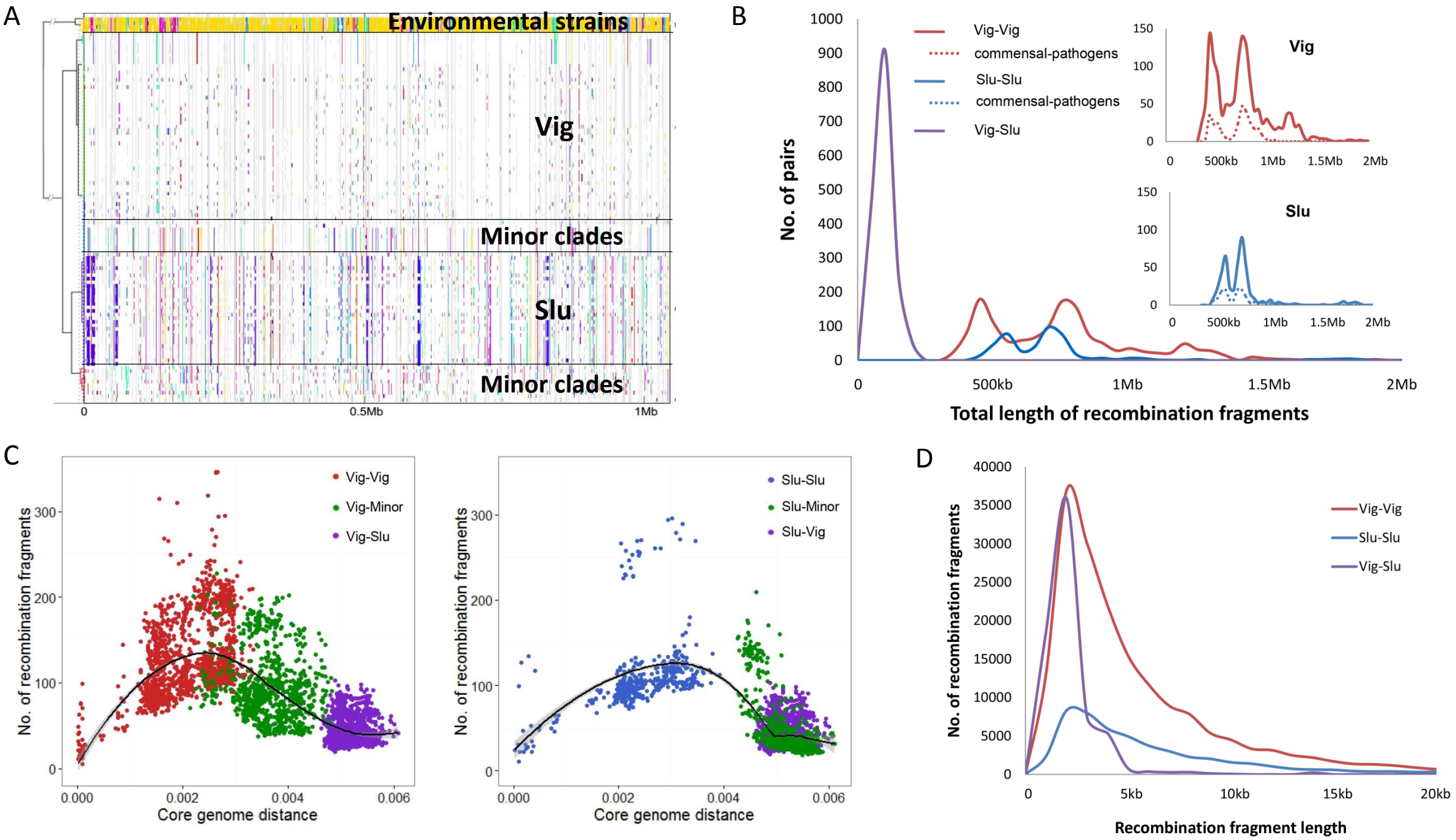

Amino acid
Carbohydrate
0.01
0.01

B


Slu vs. Slu
Vig vs. Vig
Vig vs. Slu
Frequency
800
700
600
500
400
300
200
100
0
0
0.50
1.00
1.50
2.00
Distance between homologues (%)

C


Whole Genome
GC %
51.4
51.2
51.0
50.8
50.6
50.4
**
Slu
Vig
Core Genome
GC %
53.1
53.0
52.9
52.8
52.7
**
Slu
Vig
Dispensable Genome
GC %
52.2
51.9
51.6
51.3
51.0
50.7
**
Slu
Vig


A
Environmental strains
Vig
Minor clades
Slu
Minor clades
0
0.5Mb
1Mb
B
Vig-Vig
commensal-pathogens
Slu-Slu
commensal-pathogens
Vig-Slu
No. of pairs
1000
900
800
700
600
500
400
300
200
100
0
0
500kb
1Mb
1.5Mb
2Mb
Total length of recombination fragments
Vig
150
100
50
0
0
500kb
1Mb
1.5Mb
2Mb
Slu
150
100
50
0
0
500kb
1Mb
1.5Mb
2Mb
C
Vig-Vig
Vig-Minor
Vig-Slu
No. of recombination fragments
300
200
100
0
0.000
0.002
0.004
0.006
Core genome distance
Slu-Slu
Slu-Minor
Slu-Vig
No. of recombination fragments
300
200
100
0
0.000
0.002
0.004
0.006
Core genome distance
D
No. of recombination fragments
40000
35000
30000
25000
20000
15000
10000
5000
0
0
5kb
10kb
15kb
20kb
Recombination fragment length
Vig-Vig
Slu-Slu
Vig-Slu