이학박사학위논문

대규모 유전체 분석을 통한 대장균의 유전체 다양성 및 진화 분석

Evolutionary Studies on Intra-Species Genomic Diversity of *Escherichia coli* using Large Scale Genome Analysis

2016년 8월

서울대학교대학원

생명과학부

이 기 현

이학박사학위논문

대규모 유전체 분석을 통한 대장균의 유전체 다양성 및 진화 분석

Evolutionary Studies on Intra-Species Genomic Diversity of *Escherichia coli* using Large Scale Genome Analysis

2016년 8월

서울대학교대학원

생명과학부

이 기 현

Evolutionary Studies on Intra-Species Genomic Diversity of *Escherichia coli* using Large Scale Genome Analysis

Advisor: Professor Jongsik Chun, Ph. D.

by Kihyun Lee

Submitted in Partial Fulfillment

of the Requirements for the

Degree of Doctor of Philosophy

August 2016

School of Biological Sciences

Seoul National University

대규모 유전체 분석을 통한 대장균의 유전체 다양성 및 진화 분석

지도교수 천 종 식

이 논문을 이학박사 학위논문으로 제출함

2016년 6월

서울대학교 대학원

생명과학부

이 기 현

이기현의 박사학위 논문을 인준함

2016년 8월

위 원 장 _____

부위원장 _____

위 원 _____

위 원 _____

위 원 _____

ABSTRACT

Bacterial evolution is driven by enormous genomic diversity present in the populations. Genomic diversity of a bacterial population is generated and maintained by the compounded influences of several microevolutionary mechanisms. Uniqueness of bacterial genome evolution originates from the mixture of vertical and horizontal heredity. As a result the dynamics of bacterial genomes within a species exhibits both the characters of clonal and sexual genetics, and impressively, seemingly unlimited genomic repertoire could be achieved by a single species. The course and consequences of genomic diversification within bacterial species have not been fully understood. Because of extensive genomic diversity within a species, understanding of the genomic evolution within bacterial species requires large scale exploratory and descriptive studies as well as explanatory studies based on the working hypothesis on how genomes evolve. A well-known laboratory model organism *Escherichia coli* has been shown to exploit highly diverse ecological niches in its natural population. Genomic studies of *E. coli* indeed revealed significant genome dynamism accompanied by ecological diversification. *E. coli* includes several types of pathogens that have exerted severe global burden of enteric diseases, and by that reason, whole genome surveys have been active for this species. At this point of time more than four thousands of *E. coli* genome sequences from genetically diverse strains have become available. Therefore *E. coli* constitutes an ideal model for studies of intra-specific genomic evolution of bacteria. In this thesis, multiple aspects of the genomic diversity of *E. coli* were explored and described by comparative analysis of 3,945 genome

sequences of the strains belonging to the genus *Escherichia*. In addition the roles played by distinct microevolutionary mechanisms in the shaping of current structure of genomic diversity were assessed. Lastly a broader perspective on the evolution of *E. coli* genomes was achieved by analyzing the evolutionary history of *E. coli* and its closest relatives.

Exploration of the genomic diversity of *E. coli* was conducted in 4 aspects, by analyses of pan-genome size, sequence diversity, structural diversity and phylogenetic diversity. Openness of *E. coli* pan-genome was indicated from the analysis of 3,909 *E. coli* strains. Comparison between the phylogenetic diversity and the pan-genome size estimated for randomly selected subsets of the strains showed a linear relationship between the two values. Counter-intuitively the relative ratio of pan-genome size growth over the increment of phylogenetic diversity was higher in the phylogenetic groups of *E. coli* than for the entire species. Seeking for the reason behind this trend comprised a major theoretical motivation of this thesis. Sequence diversity of *E. coli* core genes had a unimodal distribution with 1.3% as the modal value. The core gene order was unexpectedly well conserved among *E. coli* genomes and the presence of clonal frame was supported by the linkage analysis, both indicating that the core-genome of *E. coli* was highly stable. An emerging conclusion from the analysis of genomic diversity was that the paces of gene contents diversification and gene sequence diversification can be uncoupled.

Based on whole genome scale phylogenetic analysis the phylogenetic structure was clearly present among the strains of *E. coli.* The nature of given phylogenetic structuring of *E. coli* population was another major theoretical motivation of this thesis. Increased inter-SNP linkage within the phylogenetic

groups provided a clue that each phylogenetic group has relatively elevated clonality, while recombination rates in the ancestral population of *E. coli* were higher than the current rates. Assumption of clonality within phylogenetic groups could provide an explanation for the observed higher rate of within-group pan-genome growth rate per phylogenetic diversity expansion. Increased clonality is expected to result in increased efficiency of selective sweep caused by positive selection, thus resulting in the destruction and delay of sequence diversification. Inferences of recombination history in the core-genome of *E. coli* identified that 0.78% - 4.1% of the DNA segments in the core-genome has been replaced by homologous recombination. Among the extant lineages of *E. coli* the relative impact of recombination over mutation in the changes introduced to DNA sequences was distributed around 0.6 – 0.8. Relatively recent branches showed lower R/Theta than the ancestral branches, implying historical decline of recombination's influence. This direct observation of temporal decline of recombination' supported the hypothesis of *E. coli*'s shifting toward clonality.

In the pan-genome of *E. coli* the singleton genes that occurred in just a single strain of *E. coli* could be originated from recent horizontal gene transfer or recent duplication. About half of the singleton genes could not be matched to any other genes in the current prokaryotic genome database. For about 10% of the *E. coli* singleton genes, highly similar proteins were found in diverse taxonomic divisions. Most frequently the best hits resided in the close relatives of *E. coli* in the *Enterobacteriaceae* family. However, distant taxa in other phyla, especially the *Firmicutes*, contributed significant amount of best hits, implying

that those microbes share the common environmental gene pool with natural *E. coli* population.

Predominant direction of natural selection in *E. coli* genes were shown to be negative selection, which suppresses the diversification of sequences. Strength of negative selection was stronger in the core-genome in comparison to the genes with lower gene frequency. Despite that negative selection was dominant across all gene frequency spectrum, some genes exhibited dN/dS larger than 1 and seemed to be positively selected. Transposases comprised the largest proportions of positively selected genes. Multiple genes involved in flagellar biosynthesis were detected to be positively selected or have been under relaxed negative selection.

Based on the phylogenetic analysis of 21 genera in *Enterobacteriaceae* using their core-genome, the diversification within *Enterobacteriaceae* was characterized by the pattern of radiation and extensively conflicting phylogenetic signals at the basal area. Such ambiguity at deep branches were also observed for phylogenetic networks within the genus *Escherichia*. Temporally fragmented speciation might be supported by the observation. In attempt to resolve the divergence order between the species in *Escherichia*, Bayesian multi species coalescent analysis was carried out using 3 gene sets each composed of 60 core genes. The reconciled species tree and the collective graph of the coalescences estimated by the gene set re-confirmed that the divergence order between *Escherichia* spp. are ambiguous in reality. To add the geological time-scale information to the knowledge about *E. coli* evolution, a time-tree analysis was performed on the core-genome and the previously estimated divergence time of *E. coli*. By extending the previously known

divergence time between *E. coli* and *Salmonella enterica* the age of *Escherichia* was shown to be between 37.9 – 40 MYA. The age of *E. coli* was estimated to be between 16.6 – 17.7 if the clade I was excluded from *E. coli* and 25.9 – 26.9 MYA if the clade I was included in *E. coli.* The obscurity of phylogenetic scenario for the origin of *Shigella* pathogens within *E. coli* was tackled by the comparison between multigene phylogeny of *Shigella* virulence plasmids and the chromosomal phylogeny. At least five independent plasmid acquisition events had to be assumed to explain the incongruence between the two phylogenies.

According to the results obtained in this study, population genetics of *E. coli* went through a transition from relatively sexual global population to relatively clonal sub-populations. Such a transition can provide the basis for the presence of phylogenetic structure, which is not common in bacterial species. Strong clonality was shown to have negative association with the genetic diversity of species, and the slowed sequence diversification due to the reduced recombination might be the reason for increased pan-genome growth rate per phylogenetic diversity in the phylogenetic groups of *E. coli.* As shown in the example of *E. coli*, bacterial genome evolution is affected by complex interplay between evolutionary mechanisms, and moreover, can be shifted in the course of intra-specific evolution. Therefore, the nature and concept of species and speciation in bacteria could be variable from species to species, and from time to time.

Keywords: *E. coli*, Bacteria, Evolution, Genomics, Phylogenetics, Species, Pan-genome

TABLE OF CONTENTS

ABSTRACT …… i

TABLE OF CONTENTS ……..…..... vi

LIST OF FIGURES ……..…..... ix

LIST OF TABLES ……..……...….. xiv

ABBREVIATIONS ……..…..…… xv

CHAPTER 1. General introduction ……. 1

1.1. Bacterial genome evolution 2

1.2. *Escherichia coli* ……...……. 9

1.3. Purposes and organization of this study ……...……... 12

CHAPTER 2. Analysis of intra-specific genomic diversity of *E. coli* represented in the genome dataset 14

2.1. Introduction ……..... 15

2.2. Materials and methods ……..... 19

2.2.1. Newly sequenced *E. coli* genomes and the genome data obtained from public databases ……. 19

2.2.2. Taxonomic identification, annotation of protein-coding genes and clustering of orthologous proteins …….. 24

2.2.3. Pan-genome statistics ……... 26

2.2.4. Phylogenetic analysis ……... 28

2.2.5. Population structure inference using core single nucleotide polymorphisms …….. 29

2.2.6. Analysis of gene contents variation, gene order conservation and genome-wide linkage between SNP sites ……... 30

2.3. Results 32

2.3.1. Basic characterization of the genomes data …….. 32

2.3.2. Open pan-genome of *E. coli* ……. 39

2.3.3. Statistical analysis of pan-genome gene frequency distribution ….. 46

2.3.4. Evolutionary rate of pan-genome growth ……. 52

2.3.5. Phylogenetic and population genetic structure inferred from genome data 57

2.3.6. Intra-specific sequence diversity in the pan-genome of *E. coli* 64

2.3.7. Analysis of gene content variation …….. 68

2.3.8. Conservation of synteny and linkage over long distance ……. 74

2.3.9. Comparison of *E. coli* pan-genome properties and phylogenetic structure with those of other bacterial species ……..... 79

2.4. Discussion ……..…….... 86

CHAPTER 3. Characterization of microevolutionary processes that mediated genomic diversification of *E. coli* ……..... 93

3.1. Introduction …… 94

3.2. Materials and methods ……..... 97

3.2.1. Genome dataset ……..... 97

3.2.2. Analysis of homologous recombination events ……..... 98

3.2.3. Analysis of gene gain and loss history and tracking the origins of the singleton genes in *E. coli* pan-genome ……. 100

3.2.4. Analysis of dN/dS ratio …….. 102

3.3. Results …….. 103

3.3.1. Impact of homologous recombination in genomic evolution of *E. coli* …….. 103

3.3.2. Impact of gene gain and loss in the genomic evolution of *E. coli* and the origins of recently gained genes ……..... 119

3.3.3. Analysis of the signs of natural selection in the pan-genome of *E. coli* …….. 128

3.4. Discussion ……..……... 136

CHAPTER 4. Systematics study of *E. coli* and related taxa …….. 143

4.1. Introduction …….……..... 144

4.1.1. Timed history of bacterial evolution …….. 144

4.1.2. Obscurities in the systematics of *E. coli* ……..... 147

4.2. Materials and methods 149

4.2.1. Reconstruction of *Enterobacteriaceae* phylogeny …….…..... 149

4.2.2. Molecular clock analysis and species tree analysis of *Escherichia* …….. 151

4.2.3. Reconstruction of *Shigella* virulence plasmid phylogeny ……..... 153

4.2.4. Reconstruction of *rut* and *phn* operon phylogenies ……..... 155

4.3. Results …….……..... 156

4.3.1. Phylogenomic analysis of the evolutionary relationships of *Enterobacteriaceae* species 156

4.3.2. Molecular chronology of *E. coli* …….…..... 168

4.3.3. Phylogenetic scenario for *Shigella* spp. …….……. 170

4.3.4. Genes that distinguished *E. coli* from other *Escherichia* spp. …..... 175

4.4. Discussion …….…..…….. 181

CHAPTER 5. Conclusions …….…... 189

REFERENCES …….……..... 197

국문초록 …….……... 219

LIST OF FIGURES

Figure 1. Growth of the number of bacterial genome sequence data …..... 4

Figure 2. Correlation between the number of protein-coding genes and the genome size ……...…….. 7

Figure 3. Distribution of the Ortho-ANI value between the genome sequences used in this study and the genome of the type strain of *E. coli* ……. 33

Figure 4. Variations of *E. coli* genome size and the proportions of intergenic sequences along the phylogenetic tree of the strains ……. 36

Figure 5. Distribution of isolation sources data in the strains used in this study ……... 37

Figure 6. Contributions made by the genome sequences produced in this study to the availability of genome data of non-human isolates ……. 38

Figure 7. Sampling-dependent pan-genome growth curve of *E. coli*/*Shigella* strains …… 42

Figure 8. Power-law function fitted to the pan-genome growth curve of *E. coli/Shigella* strains ……... 43

Figure 9. Pan-genome growth curves estimated separately for each phylogenetic group ……. 44

Figure 10. Gene frequency distribution of the pan-genomes of *E. coli* and *Shigella* ……... 47

Figure 11. Gene frequency distribution estimated for each phylogenetic group ……... 48

Figure 12. Statistical fitting of the left end of the gene frequency distribution of *E. coli* pan-genome ……... 50

Figure 13. Relationship between pan-genome size and the phylogenetic diversity of the strains …….. 54

Figure 14. Comparison of the linear regressions obtained for the pan-genomes of different groups ……. 55

Figure 15. Negative association between the slopes of pan-genome growth per phylogenetic diversity and the total phylogenetic diversity of the group of strains ……... 56

Figure 16. Maximum-likelihood phylogenetic tree of the strains in *Escherichia* ……... 58

Figure 17. Schematic representation of the evolutionary relationships of the phylogenetic groups …….. 59

Figure 18. Subpopulation clustering and posterior probabilities of affiliation made for each strains by DAPC method …… 61

Figure 19. LD-decay plots estimated for *E. coli* and for each phylogenetic group within *E. coli* ……. 63

Figure 20. Distribution of nucleotide diversity of the genes in different gene frequency categories, from core genes to rare genes ……. 65

Figure 21. Distribution of nucleotide diversity on the chromosome of MG1655 strain ……. 67

Figure 22. Presence/absence distribution of dispensable genes among *E. coli* strains …… 69

Figure 23. Group-specific genes in *E. coli* pan-genome ……. 70

Figure 24. Distribution gene content dissimilarities within *E. coli* ……. 72

Figure 25. Relationship between the gene content dissimilarity and core-genome sequence difference …….. 73

Figure 26. Conservation of synteny within *E. coli* ……. 75

Figure 27. LD-decay within *E. coli* genomes at short and long physical distance ……... 77

Figure 28. Chromosome-wide linkage heat-map based on D' values ……. 78

Figure 29. Gene frequency distribution of diverse bacterial species …….. 81

Figure 30. Pan-genome growth curves normalized by phylogenetic diversity of the strains, for 39 diverse bacterial species ……... 82

Figure 31. Relative ratio between pan-genome growth and phylogenetic diversity of diverse bacterial species …… 83

Figure 32. Nucleotide diversity of core genes in diverse bacterial species …… 85

Figure 33. Properties of recombination events detected in the core-genome of *E. coli* …….. 105

Figure 34. Total impact of recombination on the genomes of each phylogenetic group ……... 106

Figure 35. Heat-map of the number of recombination events per gene, per donor-recipient group combination ……... 107

Figure 36. Distribution of recombination frequency per gene per strain …….. 108

Figure 37. Co-occurrence of recombination in the strains phylogenetically closely related to each other …… 110

Figure 38. Amount of gene flow per source-receiver phylogenetic groups ….. 111

Figure 39. Decline of R/theta throughout the evolutionary history of *E. coli* ... 113

Figure 40. Values of R/theta estimated in the terminal branches …….. 114

Figure 41. Prevalence of recombination in the genes compared by the gene frequency categories ……... 116

Figure 42. Proportions of recombined core genes in 39 bacterial species ……. 117

Figure 43. Relationship between the prevalence of recombination in the core-genome and the genetic diversity of species …… 118

Figure 44. Profiles of the ublast hits of the singleton genes of *E. coli* found in the genomes of *E. coli* ……... 121

Figure 45. Amino acid identity distribution between the *E. coli* singleton genes and their hits found in the genomes of other species ……... 122

Figure 46. Historical trend of gene gain and loss rates ……. 127

Figure 47. Distribution of dN/dS ratio of the genes in different gene frequency categories …… 129

Figure 48. Distributions of dN and dS in the genes belonging to 4 different gene frequency categories ……... 130

Figure 49. Within group dN/dS estimations vs. species-level dN/dS estimation ……. 135

Figure 50. Neighbor-Net phylogenetic network of *Enterobacteriacaeae* based on the core-genome ……..... 158

Figure 51. Neighbor-Joining tree of the *Enterobacteriaceae* based on the core-genome ……... 159

Figure 52. Neighbor-Net phylogenetic network of *Escherichia*, *Citrobacter*, *Salmonella* and *Enterobacter* strains ……... 160

Figure 53. Maximum credibility species tree derived from species trees inferred by multi species coalescence …..……..…….. 162

Figure 54. DensiTree graphical representation of the signals generated in multiple species trees …… 165

Figure 55. Timing of divergence between the species and clades of *Escherichia* ……. 169

Figure 56. Occurrence of the RBHs of *Shigella* VP genes in the genomes of *E. coli* and *Shigella* strains …….. 172

Figure 57. Selection of *Shigella* VP carrier strains and selection of the core genes of *Shigella* VPs ……..……. 173

Figure 58. VP core gene phylogeny vs. chromosomal SNP phylogeny …… 174

Figure 59. Multi-gene phylogeny of the *rut* operon of *E. coli* and its homologs ……. 179

Figure 60. Multi-gene phylogeny of the *phn* operon of *E. coli* and its homologs ……. 180

LIST OF TABLES

Table 1. Summary statistics of the basic properties bacterial genomes …… 6

Table 2. The strain names and BioProject accession numbers of the genome sequences produced in this study …….. 21

Table 3. Genome datasets of diverse bacterial species that were analyzed in this study ……. 23

Table 4. Power law functions obtained for the pan-genome growth curves of phylogenetic groups of *E. coli* ……... 45

Table 5. Comparison between the statistical fittings results from the gene frequency distributions of different *E. coli* phylogenetic groups ….. 51

Table 6. Recombination-hot genes of *E. coli* …… 109

Table 7. Taxonomic composition of the best hits of *E. coli* singleton genes whose homologues were found only outside of *E. coli* …… 123

Table 8. The top 25 genera that contributed most to the best hits to the *E. coli* singleton genes ……... 125

Table 9. Functions of the positively selected genes in the *E. coli* pan-genome ……. 132

Table 10. Functions of the high-frequency genes that were suspected to be under positive selection or relaxed negative selection …….. 133

Table 11. Functions of the gene sets that are diagnostic to *E. coli* …...……... 176

ABBREVIATIONS

NCBI	National center of biotechnology information
PCR	Polymerase chain-reaction
MLEE	Multi-locus enzyme electrophoresis
MLST	Multi-locus sequence typing
EHEC	Enterohemorrhagic *E. coli*
EPEC	Enteropathogenic *E. coli*
ETEC	Enterotoxigenic *E. coli*
EAEC	Enteroaggregative *E. coli*
DAEC	Diffusely adherent *E. coli*
AIEC	Adherent invasive *E. coli*
EAHEC	Enteroaggregative-hemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
VP	Virulence plasmid
TEC	Thermophilic E. coli medium
mFC	modified fecal coliform medium
SE reads	Single-end reads
PE reads	Paired-end reads
SNP	Single nucleotide polymorphism
LD	Linkage-disequilibrium
DAPC	Discriminant analysis of principal components
dN	Nonsynonymous substitution rate

dS	Synonymous substitution rate
RBH	Reciprocal best hit
Kb	Kilobase
Mb	Megabase
MCMC	Markov chain Monte Carlo
BLAST	Basic local alignment search tool

CHAPTER 1

General introduction

1.1. Bacterial genome evolution

Polymorphisms present in the hereditary DNA of organisms, the genome, provide the raw material of evolution. Understanding of the genome's evolution forms the foundation on which more practical inquiries can be made regarding the evolution of organism's phenotypes. As stated by Carl Woese, molecular sequences are "generally more revealing of evolutionary relationships" than are phenotypes, "particularly so among microorganisms" (Woese, et al. 1990). Microbial species richness on earth was estimated to reach 10^{11} - 10^{12} species and shown to be greater than those of macroorganisms (Locey and Lennon 2016). In a recently proposed view of the tree of life that outlined the diversity and evolutionary relationships of the three domains of life, lineages of microorganisms in *Archaea*, *Eukarya* and *Bacteria* comprised the majority of biodiversity (Hug, et al. 2016). It is reasonable to say that bacterial evolution is the process that generated the major part of the Earth's biodiversity. Technological advancements of the field of genome sequencing within the last decade were astonishing. As a result nowadays evolutionary analysis based on whole genome sequences is commonplace. The first bacterial genome to be completely sequenced was the genome of *Haemophilus influenza* Rd (Fleischmann, et al. 1995) and the second one was that of *Mycoplasma genitalium* G37 (Fraser, et al. 1995). Since then the number of publicly released bacterial genome sequence data kept growing in exponential rate (**Fig. 1**). At the time of writing 67,040 assembled bacterial genome sequences were available in the genome database of NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/). Discoveries

made in bacterial genomics during the last 20 years changed our view of evolution (Koonin and Wolf 2012). Among the surprising realizations brought by bacterial genomics, the most outstanding one would be the discovery of widespread horizontal gene transfer events. It was recognized that evolution of bacterial genomes does not fit into the traditional model of evolution symbolized by bifurcating lineages. Because of horizontal gene transfer events bacterial genes often have evolutionary history that is different from that of the other genes in the genome (Treangen and Rocha 2011b). Another revolutionary finding was that the genomic diversity in a single bacterial species can be unexpectedly huge (Tettelin, et al. 2005). The concept of pan-genome stemmed out from the recognition of extensive gene content variation among the strains of the same species. The two characteristics of bacterial genome evolution challenged bacterial species concept. Concept of species has not been established for prokaryotes and it is still difficult to define what bacterial species is and what drives the speciation process (Cohan and Perry 2007; Doolittle and Zhaxybayeva 2009; Fraser, et al. 2009). This study was designed to be an in-depth analysis of the genomes of a bacterial species, *Escherichia coli*, to present a comprehensive description of *E. coli* from genomics perspective. As an exemplary model organism, *E. coli* was selected because their genomic diversity and ecological diversity have been established well and rich genomic data was available for *E. coli*.

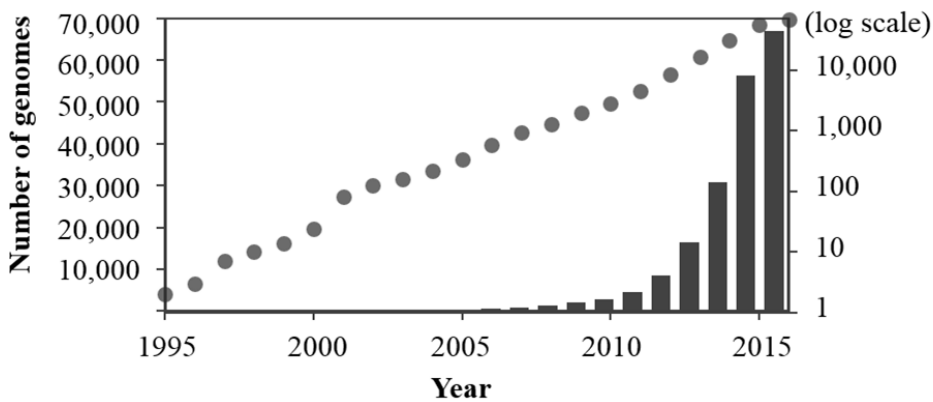

Figure 1. Growth of the number of bacterial genome sequence data. Each vertical bar represents the accumulated number of genome sequences (including draft assembly) that were available at the end of the corresponding year. To demonstrate exponential growth rate, a log-scaled secondary axis was added (circles).

General properties of bacterial genomes are different from that of more familiar eukaryotic model organisms. A general-sized bacterial genomes have smaller size than eukaryotic genomes. **Table 1** provided the summary statistics of the genome size, the number of protein-coding genes and G+C content of the completely sequenced bacterial genomes. Genome size spectrum of bacteria span two orders of magnitude, from 0.11 Mb to 15 Mb. The genome size distribution partially reflect the diversity of lifestyles of bacteria. Genome sizes tend to be extremely small in endosymbiotic bacteria that live in restricted environmental conditions and depend much of the essential functions on their host cells (McCutcheon, et al. 2009; McCutcheon and Moran 2012; Bennett and Moran 2013; Bennett, et al. 2016). In the opposite extreme of genome size, species that exhibit "social behavior" exhibit the largest genomes observed. Genomes larger than 10 Mb are observed mostly from the phylum *Actinobacteria* or the order *Myxococcales* (Bentley, et al. 2002; Han, et al. 2013). Except for such extreme cases, most of the known bacterial genomes have size of 1-10 Mb. From the summary of coding capacity of bacterial genomes, it is safe to say that bacterial genomes contain somewhere between 500 ~ 10,000 protein-coding genes. Unlike large genomes of some eukaryotic organisms, importance of non-coding fragments is relatively insignificant for bacterial genomes. Bacterial genomes are almost fully composed of protein-coding regions as demonstrated by the linear correlation between the number of protein-coding genes and the size of genomes (**Fig. 2**). The slope of linear function in the **Fig. 2** can be interpreted as 1.13 Mb increase of genome size for every thousand more genes.

Table 1. Summary statistics of the basic properties bacterial genomes.

Genome properties	Minimum	Median	Maximum
Genome size	0.112 Mb	3.70 Mb	14.8 Mb
Number of protein-coding genes	116	3,234	11,518
G+C content	13.5 %	47.5 %	74.9 %

The summary statistics were estimated based on the values observed in 5,075 completely sequenced bacterial genomes.

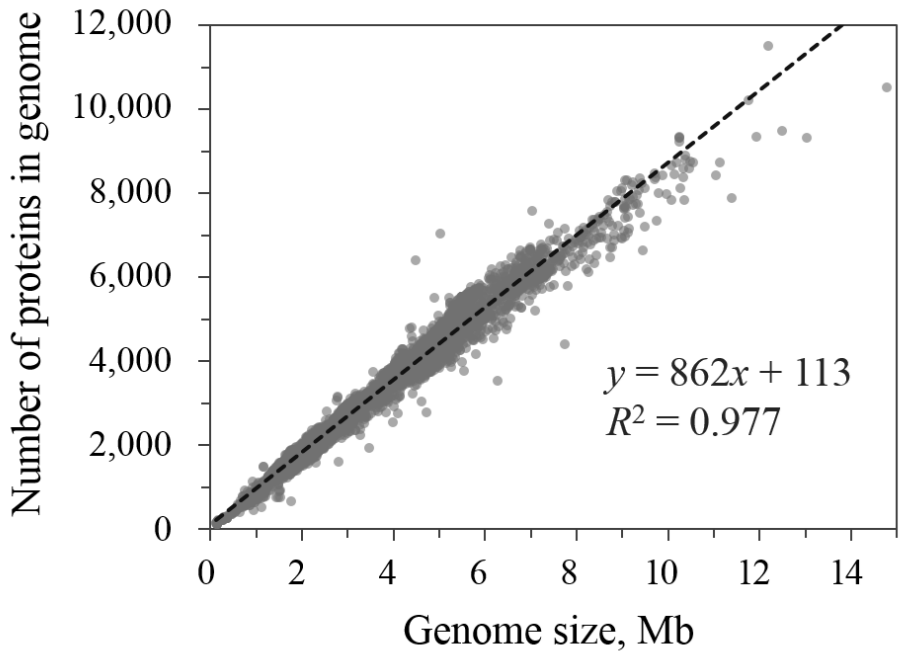

Figure 2. Correlation between the number of protein-coding genes and the genome size. Data points were derived from completed bacterial genomes. A tight linear correlation could be observed by the linear regression line (dashed line) inserted in the figure. The tight correlation indicated that the variations of intergenic sequence contents have insignificant impact to bacterial genome size.

Even with small size and compact structure, the dynamics of bacterial genomes has been shown to be unexpectedly complex. Unique and eminent features found in bacterial genomes are the mobile genetic elements. More than 10 types of mobile genetic elements have been known to exist in bacterial genomes (Koonin 2003; Brüssow, et al. 2004; Treangen and Rocha 2011a; Koonin and Wolf 2012; Darmon and Leach 2014; Soucy, et al. 2015). These genomic features are inherited from lineage to lineage within and across the taxonomic borders, to create uniqueness and complexity in the model of bacterial evolution. Diversity and prevalence of the mobile genetic elements distinguish the genomics of bacterial from the genomics of human, animals and plants.

1.2. *Escherichia coli*

Microbiologists began to recognize the presence of *E. coli* 130 years ago when a strain of the *E. coli* was cultivated by Theodor Escherich in the summer of 1884 (Escherich and Bettelheim 1988). *E. coli* has become one of the most popular laboratory model organism among bacteria. Accordingly, biochemistry, molecular biology and systems biology of *E. coli* have been studied extensively in the laboratory. Rich knowledge of subcellular machineries and molecular interactions governing the life of *E. coli* was developed as a few strains of *E. coli* were used frequently as a laboratory model. Compared to such studies, evolutionary studies of *E. coli* as a naturally occurring member of microbial ecosystems have not received as much attention. In the natural habitats the population of *E. coli* has survived and proliferated for tens of millions of years. Accordingly the current descendants of this species harbors great diversity of genomic and ecological traits. Ecological diversity of *E. coli* is manifested in the existence of numerous lifestyles. Strains of *E. coli* have been found to be gut commensals, at least 8 distinct types of enteric pathogens, extra-intestinal pathogens and free-living. Researchers have invested their efforts to explain the course of genome evolution that accompanied such phenotypic variations among *E. coli* strains. Historic contributions were made from MLEE and MLST studies (Tenaillon, et al. 2010). More recently, studies of genome sequences have driven the major efforts (Chattopadhyay, et al. 2009; Touchon, et al. 2009; Lukjancenko, et al. 2010; Leopold, et al. 2011; Luo, et al. 2011; Didelot, et al. 2012; Kaas, et al. 2012; Gordienko, et al. 2013; Bohlin, et al. 2014; Bobay, et al. 2015). In addition, molecular genotyping surveys using discriminative PCR (Unno, et al. 2010; Jang, et al. 2014), phylogenetically diagnostic PCR (Clermont, et al. 2013) and

antibiotics-resistance surveys (Amaya, et al.) performed using isolate collections have complemented the genomics researches by providing the environmental distribution of genotypes and phenotypes of *E. coli*.

E. coli have been isolated from a wide range of animals. A comprehensive survey of variety of animals indicated that *E. coli* is more frequently isolated from warm-blooded than cold-blooded animals (Gordon and Cowling 2003). *E. coli* was also reported to be isolated from the environments outside of the host, such as soil, freshwater and sediments (Goto and Yan 2011). Some of the environmental isolates or cold-blooded animal isolates should be considered as anthropogenic effect, but studies have shown that *E. coli* strains can genuinely inhabit the out-of-host environments (Byappanahalli, et al. 2003; Ishii, et al. 2006). Inside the host animal, the habitat of *E. coli* is divided into the primary habitat, which is the gastro-intestinal tract and the secondary habitat, the extra-intestinal body sites. Apart from such a diverse habitats of *E. coli*, phylogenetic diversity within *E. coli* has been known for long time. Early studies based on MLEE already identified the existence of sub-species level structures of *E. coli* (Guttman and Dykhuizen 1994; Tenaillon, et al. 2010). Four traditionally recognized groups were defined from earlier researches and each of them was designated as group A, B1, B2 and D. Subsequently additional phylogenetic groups were recognized by the studies using MLST schemes (Tenaillon, et al. 2010). Currently there are seven recognized phylogenetic groups: A, B1, B2, C, D, E and F. While the existence of the phylogenetic groups have been well established, ecological differences of the phylogenetic groups remained largely unknown. Some studies have addressed whether or not the ecological and evolutionary divergence between the groups exist (Gordon and Cowling 2003;

Escobar-Páramo, et al. 2006; Leopold, et al. 2011; Didelot, et al. 2012; de Muinck, et al. 2013) but genome-scale studies had critical limitations because of the lack of coverage on the diversity within *E. coli.*

Phenotypic and ecological diversity of *E. coli* is also manifested in the presence of diverse lifestyles, especially in terms of commensals and pathogens. Presence of numerous different pathogenicity mechanisms made *E. coli* an attractive model species for the studies of genomic evolution that drive the evolution of pathogenicity. Based on the pathogenicity *E. coli* strains can be divided into three broad categories: commensal strains, intestinal pathogens and extra-intestinal pathogens. Virulence of pathogenic *E. coli* is achieved by the virulence factors encoded in the dispensable genes. The intestinal pathogens are further divided based on distinctive mechanisms of pathogenesis. Currently recognized enteric pathotypes are EHEC, EPEC, ETEC, EAEC, DAEC, AIEC, EAHEC and EIEC. Extra-intestinal pathogens are further divided into EnPEC, NMEC and UPEC. In addition to the above-mentioned pathotypes, *E. coli* harbors the genus *Shigella*, the strains that causes Shigellosis disease and bacterial dysentery. This unusual cases of classification was made because the strains of *Shigella* exhibit distinct phenotypic characters. The genus *Shigella* was proposed in 1949 and currently there are four species of *Shigella*, Lack of motility and inability to ferment lactose distinguishes the strains of *Shigella* from *E. coli*. Unlike other *E. coli* pathotypes, the members of *Shigella* are obligate pathogens. Multiple phylogenetic studies have pointed out that the *Shigella* strains are actually specific lineages within the species *E. coli* (Pupo, et al. 2000; Sims and Kim 2011; Sahl, et al. 2015). A general consensus made from the

previous studies was that *Shigella* emerged in multiple occasions, and parallel convergent evolution made their phenotypic and genomic coherence.

1.3. Purposes and organization of this study

This study was aimed to exemplify the extent of genomic diversity within a bacterial species and explain how genomic diversify could be shaped by unique evolutionary processes of bacterial genome evolution. Contents of the thesis was divided into three chapters. Chapter 2 focused on the description of the extent of genomic diversity. Genomic diversity will be assessed from multiple angles. Properties of *E. coli* pan-genome will be analyzed to evaluate the diversity of gene repertoire. Sequence diversity will be estimated throughout the pan-genome and the conservation of gene order and linkage will be tested. Phylogenetic diversity and subspecies level structure inside the species will be tested. Chapter 3 was devoted to explain the microevolutionary processes that generated the genomic diversity of *E. coli* observed in the chapter 2. Contributions of recombination, gene gain and loss, and horizontally transferred genes were estimated to that end. In addition the role of natural selection in the shaping of genetic diversity was tested. Chapter 4 was added to provide broad perspective to the genomics of *E. coli*. Evolution of the species *E coli* was analyzed in longer time-scale, in the context of the family *Enterobacteriaceae* and the genus *Escherichia*. Phylogenetic scenario for the rise of *Shigella* spp. within *E. coli* was investigated too. Finally, the genes that distinguish *E. coli* from the other *Escherichia* species were identified and described in depth.

CHAPTER 2

Analysis of intra-specific genomic diversity of *E. coli* represented in the genome dataset

2.1. Introduction

After 21 years since the publication of a the genome sequence of *H. influenza* Rd strain (Fleischmann, et al. 1995), whole genome sequencing and analysis has become the common method of choice for determination of genetic traits of bacterial strains. Accumulated number of bacterial genome sequence data have been increasing each year at accelerating pace (Land, et al. 2015). As a result, taxonomic coverage of genome-sequenced species is expanding rapidly (Chun and Rainey 2014). For species that received major research interest, the number of published genome sequences now exceed a thousand. Species under such hot attention include *E. coli*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Salmonella enterica*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii*. Availability of the genome sequences of multiple different strains of the same bacterial species provides a unique opportunity to explore the intra-species genetic diversity of bacteria at whole genome scale. From the earliest comparative genomics studies it became evident that genomic variations within single bacterial species consist not only of the polymorphisms observed at shared loci and that significant amount of genomic loci that are present in a strain are absent in another. For example, the earliest comparative genomics study of *E. coli* were performed using the K-12 strain MG1655 and the O157:H7 strain Sakai and reported that 567 and 1,632 protein-coding genes were unique to each strain, respectively (Hayashi, et al. 2001). In that article, presence of strain-specific loci were conceptually explained by the model of common genomic ‘backbone’ and strain-specific ‘loops’.

In the seminal research article published in 2005, the result from analysis of 8 representative genomes of *Streptococcus agalactiae* were reported (Tettelin, et al. 2005). The concept of pan-genome was proposed based on the realization that the gene repertoire of bacterial species cannot be defined by the reference genome. At the same time, open and closed nature of the pan-genomes of different species were discussed in that article. Characterization of pan-genomes often paid attention to the determination of "open" or "closed" nature of each pan-genome. Openness of *E. coli* pan-genome has been suggested in multiple studies (Rasko, et al. 2008; Touchon, et al. 2009; Didelot, et al. 2012; Gordienko, et al. 2013). Actually, in previous studies of a variety of bacterial species the majority of the species were characterized to have "open pan-genome" (Park, et al. 2012; Zhang and Sievert 2014). A few cases of "closed pan-genome" were observed for *S. aureus, Streptococcus pyogenes, Ureaplasma urealyticum* and *Bacillus anthracis* based on power exponent criterion that was first suggested by Tettelin's paper (Tettelin, et al. 2008). High plasticity of bacterial genomes and immeasurable size of bacterial population make it not feasible to collect the entire global gene repertoire of a species. Previous studies of *E. coli* genomes all have concluded that the pan-genome of *E. coli* is open (Rasko, et al. 2008; Touchon, et al. 2009; Lukjancenko, et al. 2010; Kaas, et al. 2012; Gordienko, et al. 2013; Sahl, et al. 2014). The number of strains compared in the previous studies are 18 (Rasko, et al. 2008), 20 (Touchon, et al. 2009), 64 (Lukjancenko, et al. 2010), 186 (Kaas, et al. 2012), 32 (Gordienko, et al. 2013) and 100 (Sahl, et al. 2014). The dataset of 3,909 strains used in this study would provide an unprecedentedly comprehensive characterization of *E. coli* pan-genome.

Statistical properties of pan-genomes are often characterized by the shape of their gene frequency distribution. Gene frequency distributions could be affected by biological nature of the organisms such as their innate rate of gene fragment uptakes, population evolutionary parameters such as the rate at which newly acquired genes are lost or fixed, or ecological parameters like the fitness effects of gene acquisition in different strains. It can be expected that for a group of organisms in which the fitness effect for the presence/absence of the genes are invariable among the individuals (i.e. under homogeneous ecological conditions), gene frequency distribution of the population would be shifted toward right side of the spectrum. To compare gene frequency distributions of different group of organisms, model-based analyses have an advantage over visual inspection since the model-based analyses offer quantitative comparison of the estimated parameters. A previous study of *E. coli* pan-genome using 48 complete genomes suggested that the gene frequency distribution can be best approximated by the bi-component model, as the sum of two power functions (Gordienko, et al. 2013).

Pan-genome is an outstanding conceptual model that can describe the gene repertoire diversity harbored by bacterial species. Other than the diversity of gene repertoire, genomic diversity of bacterial species contain multiple additional components. Another major component is the polymorphisms and diversity of the sequences at the orthologous loci. This aspect of genomic diversity is more familiar to the classic concept of genetic diversity. The third component is the structural variations, including the differences caused by inversion, translocation, deletion, duplication or insertion of the sequences. In addition to the 3 components of genomic diversity, phylogenetic diversity is also tightly connected with the patterns of

genomic diversity. The patterns and structure in genomic diversity of a species can be described by population structure and phylogenetic structure. Bacterial species have been shown to have diverse population structures, ranging from panmictic to clonal and most bacterial species are in the middle of the spectrum (Joseph, et al. 2011; Underwood, et al. 2013). Studies of intra-species genetic structure of bacteria can be performed in the scheme of phylogenetics as well as population genetics. In this chapter, to explore various aspects of the genomic diversity of *E. coli*, 3 general topics were addressed. In the first part, the properties of *E. coli* pan-genome was examined. In the second part, phylogenetic structure among the strains was be investigated. In the third part, the level of sequence diversity and structural variations was estimated. Finally, the results obtained for *E. coli* was compared with diverse other bacterial species to evaluate the generality of the results.

2.2. Materials and methods

2.2.1. Newly sequenced *E. coli* genomes and the genome data obtained from public databases

Sampling of the feces of various domestic animals and the surface water of urban rivers was carried out in the area of Seoul and Gyeong-gi province of Korea. From the samples, strains of *E. coli* were isolated using mTEC (manually prepared), mFC (Kisan Biotech) and MacConkey (Difco) culture media. Serially diluted samples were spread on the agar plates and incubated at 37°C. Colonies were then routinely cultured on Nutrient Agar (Difco) plates. Taxonomic identification of the strains was carried out by analyzing their 16S rRNA gene sequences. Other than the strains isolated from Gyeong-gi porovince, 51 *E. coli* strains were provided by professor Hor-gil Hur of Gwangju Institute of Science and Technology. The strains provided by professor Hur included 8 strains isolated from cow feces, 8 strains isolated from chicken feces, 5 strains isolated from pig feces, and 30 strains isolated from river water all of which was isolated in the Jeolla province. Genomic DNA of the strains were purified using the standard bacterial protocol of Wizard Genomic DNA Purification Kit (Promega). For preparation of Illumina sequencing library, the standard protocol and the kit contents of TruSeq DNA PCR-Free Library Preparation Kit (Illumina) or Nextera DNA Sample Preparation Kit (Illumina) was used to generate whole genome shotgun libraries. For 41 strains, paired-end 300 bp reads were generated. For another 10 strains, paired-end 250 bp reads were generated. For the rest of the strains, paired-end or single-end 150 bp reads of Illumina platform

were generated. Specifically for 7 strains that were sequenced by 150 bp reads, additional 8 kb paired-end reads were generated by GS FLX Titanium platform (Roche) to aid assembly. Assembly of 150 bp SE, 150 bp PE, 250 bp PE reads were conducted on CLC Genomics Workbench 5.1 (CLC Genomics). Assembly of hybrid approach by combined usage of Roche GS assembler 2.6 and CLC Genomics Workbench 5.1. Assembly of 300 bp PE reads were conducted using a5-miseq pipeline of a5 assembler (Coil, et al. 2015). The resulting draft genome sequence assemblies were submitted to NCBI and EBI (Accession numbers are provided in **Table 2**).

Genome sequences of *Escherichia* strains analyzed in this study included 3,883 genome sequences obtained from the Genomes Resources of NCBI (http://www.ncbi.nlm.nih.gov/genomes/). In addition to *E. coli*, genome sequences of diverse other bacterial species were analyzed to be compared with the results obtained from *E. coli*. The list of species analyzed in this study using the data obtained from NCBI Genomes Resources are listed in the **Table 3**.

Table 2. The strain names and BioProject accession numbers of the genome sequences produced in this study.

Index	Strain name	BioProject accession
1	*E. coli* 12-GJ420	PRJEB10199
2	*E. coli* 12-3622	PRJEB10198
3	*E. coli* 11-4311	PRJEB10197
4	*E. coli* 11-3539	PRJEB10196
5	*E. coli* 11-3518	PRJEB10195
6	*E. coli* 11-2915	PRJEB10194
7	*E. coli* 11-2744	PRJEB10193
8	*E. coli* 11-2338	PRJEB10192
9	*E. coli* 10-GJ433	PRJEB10191
10	*E. coli* 10-2342	PRJEB10190
11	*E. coli* 08-GJ444	PRJEB10189
12	*E. coli* 08-4349	PRJEB10188
13	*E. coli* 08-3512	PRJEB10187
14	*E. coli* 08-2956	PRJEB10186
15	*E. coli* 08-2357	PRJEB10185
16	*E. coli* 08-2304	PRJEB10184
17	*E. coli* 07-4360	PRJEB10183
18	*E. coli* 07-4352	PRJEB10182
19	*E. coli* 07-4338	PRJEB10181
20	*E. coli* 07-3550	PRJEB10180
21	*E. coli* 07-3502	PRJEB10179
22	*E. coli* 07-2726	PRJEB10178
23	*E. coli* 07-2709	PRJEB10177
24	*E. coli* 06-3644	PRJEB10176
25	*E. coli* 06-2710	PRJEB10175
26	*E. coli* 05-3501	PRJEB10174
27	*E. coli* 05-2753	PRJEB10173
28	*E. coli* 04-4351	PRJEB10172
29	*E. coli* 04-2751	PRJEB10171
30	*E. coli* 04-2709	PRJEB10170
31	*E. coli* E955	PRJEB10169
32	*E. coli* E898	PRJEB10168
33	*E. coli* E888	PRJEB10167
34	*E. coli* E834	PRJEB10166
35	*E. coli* E831	PRJEB10165
36	*E. coli* E830	PRJEB10164
37	*E. coli* E822	PRJEB10163
38	*E. coli* E819	PRJEB10162

Table 2. Continued.

Index	Strain	BioProject accession
39	*E. coli* E229	PRJEB10161
40	*E. coli* E202	PRJEB10160
41	*E. coli* E30	PRJEB10159
42	*E. coli* U59	PRJEB10158
43	*E. coli* O76	PRJEB10157
44	*E. coli* O74	PRJEB10156
45	*E. coli* B66	PRJEB10155
46	*E. coli* U09	PRJEB10154
47	*E. coli* AM21	PRJEB10153
48	*E. coli* AM09	PRJEB10152
49	*E. coli* AK73	PRJEB10151
50	*E. coli* AK83	PRJEB10150
51	*E. coli* D03	PRJEB10149
52	*E. coli* D11	PRJEB10148
53	*E. coli* AJ28	PRJEB10147
54	*E. coli* B31	PRJEB10146
55	*E. coli* AD27	PRJEB10145
56	*E. coli* AF83	PRJEB10144
57	*E. coli* D88	PRJEB10142
58	*E. coli* P816	PRJEB10141
59	*E. coli* AD30	PRJEB9743
60	*E. coli* AD30	PRJNA157425
61	*E. coli* AI27	PRJNA89369
62	*E. coli* W26	PRJNA88641
63	*E. coli* AA86	PRJNA65321

Table 3. Genome datasets of diverse bacterial species that were analyzed in this study.

Species	Number of genomes analyzed	Outgroup used in phylogenetic tree
Mycobacterium abscessus	92	*M. tuberculosis*
Mycobacterium tuberculosis	102	*M. abscessus*
Yersinia enterocolitica	121	*Y. pestis*
Yersinia pestis	117	*Y. enterocolitica*
Yersinia pseudotuberculosis	56	*Y. pestis*
Burkholderia mallei	122	*B. thailendensis*
Burkholderia cenocepacia	30	*B. cepacia*
Bacillus anthracis	78	*B. cereus*
Bacillus cereus	122	*B. subtilis*
Bacillus subtilis	74	*B. velezensis*
Neisseria gonorrhoeae	103	*N. meningitidis*
Neisseria meningitidis	105	*N. gonorrhoeae*
Haemophilus influenzae	77	*H. parainfluenzae*
Campylobacter coli	99	*C. jejuni*
Campylobacter jejuni	111	*C. coli*
Staphylococcus epidermidis	86	*S. capitis*
Staphylococcus haemolyticus	68	*S. hominis*
Streptococcus agalactiae	107	*S. uberis*
Streptococcus mutans	124	*S. ratti*
Helicobacter pylori	106	*H. cetorum*
Oenococcus oeni	57	*O. kitaharae*
Chlamydia trachomatis	68	*C. muridarum*
Propionibacterium acnes	104	*P. acidipropionici*
Brucella abortus	91	*B. melitensis*
Vibrio cholerae	109	*V. mimicus*
Vibrio parahaemolyticus	106	*V. harvey*
Pseudomonas syringae	95	*P. fluorescence*
Legionella pneumophila	52	*L. fallonii*
Listeria monocytogenes	99	*L. innocua*
Bordetella pertussis	93	*B. avium*
Leptospira interrogans	67	*L. kirschneri*
Enterobacter aerogenes	74	*E. cloacae*
Enterobacter cloacae	99	*E. aerogenes*
Salmonella enterica	152	*E. coli*
Bacteroides fragilis	87	JH815484_s
Clostridium difficile	126	*C. sorderllii*
Enterococcus faecalis	124	*E. caccae*
Enterococcus faecium	112	*E. durans*

To avoid confusion between multiple genera that starts with the same alphabet, scientific names of the species were used without abbreviation in this table.

2.2.2. Taxonomic identification, annotation of protein-coding genes and clustering of orthologous proteins

Correctness of the taxonomic identification of the strains used in this study was tested by whole genome orthologous average nucleotide identity (OrthoANI) values between the strains (Lee, et al. 2016). ANI dendrogram was inspected at family level to identify the strains that were misplaced. In addition, the core-genome phylogenetic tree of the strains assigned as *Escherichia* was reconstructed inspected to detect misidentified strains. Genome assemblies were excluded if the genome size was abnormally small or large, or if contamination was detected based on the 16S rRNA gene fragments extracted from the contigs. Genomes sequenced in this study as well as those obtained from the public database were annotated (re-annotated) using the same procedure. Protein-coding regions were identified using Prodigal version 2.6.1. Functional annotation was attached by the database search against KEGG (Kanehisa, et al. 2016), eggNOG (Huerta-Cepas, et al. 2016) and SEED subsystems (Overbeek, et al. 2014). Database search was performed by amino acid sequence searching by ublast program of Usearch 8 (Edgar 2010). To exclude the spuriously short alignments from the annotation, minimum query coverage was set to 0.8, the minimum identity was set to 0.8, and e-value cutoff was set to 10^{-15}.

The pan-genome of *E. coli* was reconstructed by clustering of orthologous genes. Commonly used algorithms such as Ortho-MCL require all-against-all BLAST search and the computational load increase exponentially as the number of proteins increases (Chen, et al. 2007). To avoid the computational load required for exhaustive all-against-all search of homologous proteins, an agglomerative method

for pan-genome reconstruction was designed and applied. The method begins by manually defining input order of the input genomes. Input order was manually organized so that complete genomes are treated first and draft assemblies are treated later. The initial pan-genome is defined as the set of protein-coding genes present in the first genome. Then, for each round of comparison between the pan-genome and the next input genome, the pan-genome was iteratively updated, until the last genome. In each round, the proteome of the new input genome was compared with the proteins contained in the latest pan-genome. Between the two sets of proteins, the pairs of proteins that matched to each other by reciprocal best hit were determined. Genes in the input genome that had no RBH relationship with the strains processed in earlier rounds were added to pan-genome. RBH relationships detected during the process were recorded in order to construct the presence/absence profile of each orthologous gene cluster in the strains.

2.2.3. Pan-genome statistics

Pan-genome growth curve was generated by iterative subsampling of the presence/absence profile of orthologous gene clusters in the strains. When the number of strain in the dataset was *N*, for each sample size *n* between 1 and *N* (1 < *n* < N), random selection of *n* strains was performed for 500 times. For each *n*, the pan-genome size was counted, resulting in 500 independently counted pan-genome size values. Median of the 500 values and the 95% confidence intervals calculated based on the assumption of normal distribution were used in the plotting. At the same time, pan-genome growth curve was also generated and analyzed using PanGP (Zhao, et al. 2014). PanGP was used because the software provided a power-law-based statistical fitting of the pan-genome growth curve. Power-law based regression is a widely used statistical approach to test pan-genome openness (Tettelin, et al. 2008). In the test, the sampling-dependent pan-genome growth curve is first fitted to the following power function: $P = A\ x^B + C$. In the function *P* is the size of the pan-genome (number of genes in the pan-genome) and *x* is the number of strains sampled, to model the growth curve of the pan-genome size. Depending on the value of *B*, the pan-genome is considered to be open when the exponent *B* is in between 0 and 1. If $B < 0$, *P* would converge to a certain value as *x* increases and the pan-genome should be considered as a closed one (Tettelin, et al. 2008).

To normalize the pan-genome growth curve by the phylogenetic diversity present in the set of strains, a new statistical method was designed. In the new method, the phylogenetic tree that contains all strains in the dataset should be provided. In all analysis performed in this study, the input phylogenetic tree was inferred based on

the concatenate alignment of core genes of the taxa being analyzed. Phylogenetic tree based on concatenated alignment was used because the resulting branch length then would mean the divergence at core gene sequences. If phylogeny based on other data (such as phenetic distances) was used, the resulting branch length would have no significant meaning. In the method, for each random subsampling of strains the node corresponding to the most common ancestor of the selected strains was defined first. Then, the branches that are in the shortest path from the common ancestor to the selected strains (which are terminal leaves) were marked. The sum of branch length of the marked branches were calculated in non-redundant way. The resulting sum value corresponded to the phylogenetic diversity of the strains, as measured by the substitution per site across the core-genome. When the pan-genome size and the phylogenetic diversity measured in each repetitive random sampling were plotted against each other, the phylogenetically normalized version of pan-genome growth curve was obtained. Since the plot was found to be linear in many cases, linear regression test was performed to estimate the slope of the pan-genome growth.

2.2.4. Phylogenetic analysis

Nucleotide sequences in the same orthologous gene clusters were aligned using the following steps. First, each nucleotide sequence was translated to amino acid sequence. Then, the identical amino acid sequences were merged so that non-redundant set of amino acid sequences were obtained. Computational time was greatly improved when the removal of redundant sequences was performed by this step. Then multiple sequence alignment algorithm of MAFFT (Katoh and Standley 2013) was run with the parameter of '–retree' set to 2. Finally, the alignment of nucleotide sequences were generated by moving the position of the codons according to the positions of the corresponding amino acids in the protein sequence alignments. Before the reconstruction of phylogenetic trees, recombinant genes were detected and filtered out. For that purpose, the orthologous gene families that underwent homologous recombination were detected by PhiPack package (Bruen 2006). Core genes whose alignment received a p-value smaller than 0.05 by the Pairwise Homoplasy Index (PHI) test were assumed to have experienced recombination event and were excluded from the phylogenetic analysis. Next, the alignments of the non-recombinant core genes were inspected for their gap contents. Any alignment that contained too much gapped area (the cut-off varied between datasets to balance the quality of alignments and the quantity of genes used in the analysis, and usually set as 5% of the total alignment area) or contained an entry that had too much gaps (usually set as 10% of the alignment length) were excluded from the downstream analysis. The remaining core gene alignments were concatenated into a supermatrix. The supermatrix was used to run FastTree version 2.1.3 (Price, et al. 2010) to infer the phylogenetic tree using approximate maximum-likelihood method.

2.2.5. Population structure inference using core single nucleotide polymorphisms

Single nucleotide polymorphism (SNP) sites for which the allele state can be determined for most of the strains were identified by 1 to 1 alignment of the genomes to the reference genome of MG1655, using the programs in MUMmer 3.23 (Kurtz, et al. 2004). For each non-reference genome, nucmer program was run with '—mum' option to identify only the regions that can be mapped uniquely both in the query and the reference. From the output delta file, polymorphic sites was detected by show-snps program with option '-C' to exclude the SNP calls in the ambiguously aligned regions. SNPs discovered for non-reference strain were compiled into a single matrix that contained the entire reference coordinates and the SNP calls made in all strains. Core-SNP sites were determined from the matrix by 3 cutoffs. First, the sites were filtered out when the proportion of the strains that were un-aligned to that site were larger than a given cutoff. The cutoff value of 0.01 was used. Then the sites that resulted in a minor-allele frequency lower than a given cutoff were filtered out. The cutoff value of 0.01 was used in this step. Invariant sites were naturally removed during that step. And optionally, the sites that have 3 or 4 variants were filtered out to result in bi-allelic SNP set. For analysis of *E. coli* population structure, only bi-allelic SNP sites were used. Using the core bi-allelic SNPs, discriminant analysis of principal components (DAPC) algorithm implemented in the *R* package 'adegenet' was used to test the presence of population structure using *E. coli* core-SNP data (Jombart, et al. 2010). The algorithm required an assumption on the number of subpopulations present in the dataset. Changing the parameter for presumed number of subpopulations from 2 to 11, DAPC analysis was repeatedly run. The posterior probabilities of the cluster membership of each strain to each cluster was used to identify admixed strains.

2.2.6. Analysis of gene contents variation, gene order conservation and genome-wide linkage between SNP sites

As an indicator of intra-specific genetic diversity at each pan-genome locus, the nucleotide diversity statistic (π) was calculated for each DNA sequence alignment. Nucleotide diversity was defined as the average number of nucleotide differences per site between randomly chosen pair of strains. Calculation of nucleotide diversity was performed by running PhiPack (Bruen 2006) because the tool estimates π in the course of detecting the recombination events. Degree of gene content divergence was calculated for each pair of strains, using the Bray-Curtis dissimilarity index. The Bray-Curtis dissimilarity index for strain *i* and *j* was defined by the equation $BCij = 1 - 2Cij\ /\ (Si + Sj)$, where C_{ij} is the number of genes commonly present in the two strains, S_i and S_j are the total number of genes in the strain *i* and *j* respectively. Based on the pan-genome gene presence/absence matrix, group-specific genes were determined for each phylogenetic group. A gene was selected as a group-specific gene if the gene was present in >90% of the strains of the target group and was absent in >90% of the strains that did not belong to the target group.

Synteny of core genes of *E. coli* was visually inspected using a simple plotting method. In this analysis, only completely assembled genomes were used because the draft genomes contained incomplete information especially regarding the gene order. For each strain, the chromosomal order of core genes was determined based on their coordinates. Order ranks in the reference strain MG1655 was

compared against the order ranks found in the other strains. In the resulting plots, a diagonal line would be generated for a pair of strains that shared the same gene order.

Linkage analysis was performed to detect the clonal frame among the core-SNPs and to inspect decay of linkage according to increase of inter-marker physical distance. To calculate pairwise linkage disequilibrium (LD) for all pair of core-SNP sites, Haploview (Barrett, et al. 2005) program was run. Since all-vs-all calculation of 44,933 core-SNP sites will result in too large number of pairwise estimations and the majority of the values would be redundant for spatially clustered SNP sites, only a subset of SNP sites were used by randomly picking a single SNP site for every 100 bp interval. Random sub-sampling of SNP sites were replicated 3 times. Two different statistic measures of LD were used, R^2 value and *D*' value. To simulate the haploid genetics, chromosome parameter was set to be 'Y', to pretend that the SNP data is from the Y chromosome. LD values and the physical distance between the pair of SNP sites were combined to generate a plot of LD decay over physical distance. Analysis of LD decay plot was repeated for the SNP matrix of each phylogenetic group to provide the comparison between the intensity of recombination at species level and sub-population level. Beside the LD decay analysis, chromosome-wide heat-map of LD was generated using R statistical analysis platform. The heat-map was intended to visualize the presence of clonal frame which by definition is a persistence of long distance LD throughout the chromosome.

2.3. Results

2.3.1. Basic characterization of the genomes data

Taxonomic identity of the genome sequence data used in the *E. coli* pan-genome analysis was justified by their relatedness measured by Ortho-ANI statistic and confirmed by the results of phylogenetic analysis of their conserved sequences. For each of the 3,909 strains that were classified as *E. coli* or *Shigella* in this study, the genomic relatedness with the *E. coli* type strain DSM 30083 was measured by Ortho-ANI statistic. The resulting Ortho-ANI values were confined in the range 96.35-100% (**Fig. 3**). In the previous studies, a general boundary between prokaryotic species corresponded to the ANI (and variations thereof) value of 95-96% (Richter and Rosselló-Móra 2009; Kim, et al. 2014) and 96.5% (Varghese, et al. 2015). Among 3,909 strains that were classified as *E. coli/Shigella* only 6 strains had Ortho-ANI value between 96.35% and 96%. On the other hand, the genomes that were classified as *Escherichia* spp. other than *E. coli* showed Ortho-ANI value ranging from 89.1% to 95.0%. Therefore, taxonomic identity of the genomes used in this study is not in conflict with generally accepted guidelines for the identification of prokaryotic species. The second evidence for the correctness of taxonomic identity of the strains used in the study came from the later analysis of phylogenetic tree of the conserved core-genome loci. In the phylogenetic tree, the presence of outlier strains was not detected.

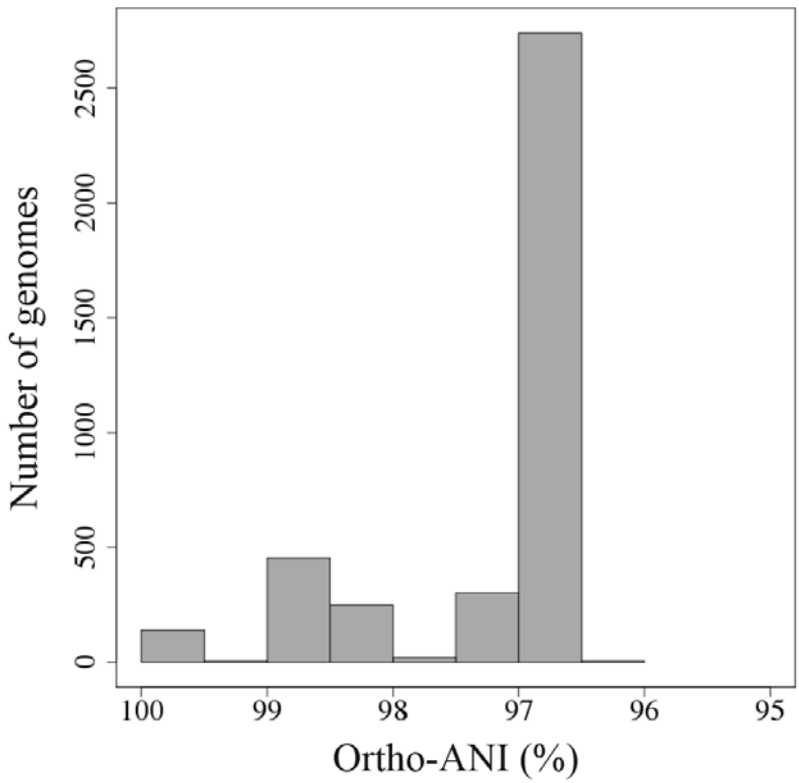

Figure 3. Distribution of the Ortho-ANI value between the genome sequences used in this study and the genome of the type strain of *E. coli.* The type strain of *E. coli* is DSM 30083. The lowest Ortho-ANI value was 96.35%. The most commonly accepted ANI cutoff for species boundary is 95-96%. The distribution of Ortho-ANI value presented in this figure demonstrated that taxonomic identity of the genome sequences used in this study could be trusted.

Basic statistics including the size of genome, the number of genes encoded in the genome, the G+C content and the proportions of intergenic sequences were characterized for each genome data. To summarize briefly, 95% of the genomes fell into the range described in the following description: the genome size was 4.56 Mb - 5.59 Mb; the proportions of intergenic sequences was 11.37% - 15.08% of the genome size; the number of protein-coding genes in the genome was 4,323 - 5,577; and the G+C content was 49.63% - 50.92%. Phyletic distribution of the genome size and the composition of intergenic regions were investigated to check if there was association between with these properties and the phylogeny. From the distribution of genome size and the proportions of intergenic sequences, a general differentiation between phylogenetic clades was not present (**Fig. 4**). Nonetheless two specific and known trends were visible in the plot: (i) reduced genome sizes of *Shigella* clades and (ii) expanded genomes of group E strains.

For genome sequence data obtained from NCBI database, we collected the metadata about isolation source, clinical information and geographical origin. After all, the host/source information for 56.1% of 3,946 strains were collected. For body site of origin, it was able to identify the information for 1,863 strains. Still many strains remained uncharacterized about identified about their relevant information. Using the collected isolation source data, host distribution and body site distribution were inspected on phylogenetic tree of the strains (**Fig. 5**). None of the host type could be specifically associated with the phylogenetic clades. Preferential occurrence of food isolates in the group E was possibly due to the fact that many food isolates in the O157:H7 clone of group E received the interest from genomics community because of the foodborne diarrheal outbreaks. In the *S. sonnei* clade,

which is known as human-specific pathogens, non-human hosts was not detected except for a single cow isolate. Body site distribution of the isolates indicated firstly that the gut isolates had high prevalence in all phylogenetic groups. Urine isolates were distributed in B2, D, F, B1 and C but not detected in *Shigella* clades, and only 1 urine isolate was found in Group E. Distributions of blood isolates and other extra-intestinal isolates were similar with that of urine isolates. Except for a general trend that extra-intestinal isolates were absent in group E and *Shigella* clades, body sites distribution was non-specific among the phylogenetic groups.

From the isolation source data collected it was obvious that although genome sequences of thousands of *E. coli* strains have been produced to date, the genome sequencing efforts have been biased toward human clinical isolates. As an effort to counter balance this bias and introduce more variety of the habitats of *E. coli* into the genome data, we presented here 62 newly sequenced genomes of non-human *E. coli* isolates. Behind this, 110 samples of feces and river waters. From the samples were collected and 969 *E. coli* isolates were collected from the samples. The contribution made by the 62 genome sequences to the abundance of the genome data of non-human *E. coli* isolates was summarized in **Fig. 6**.

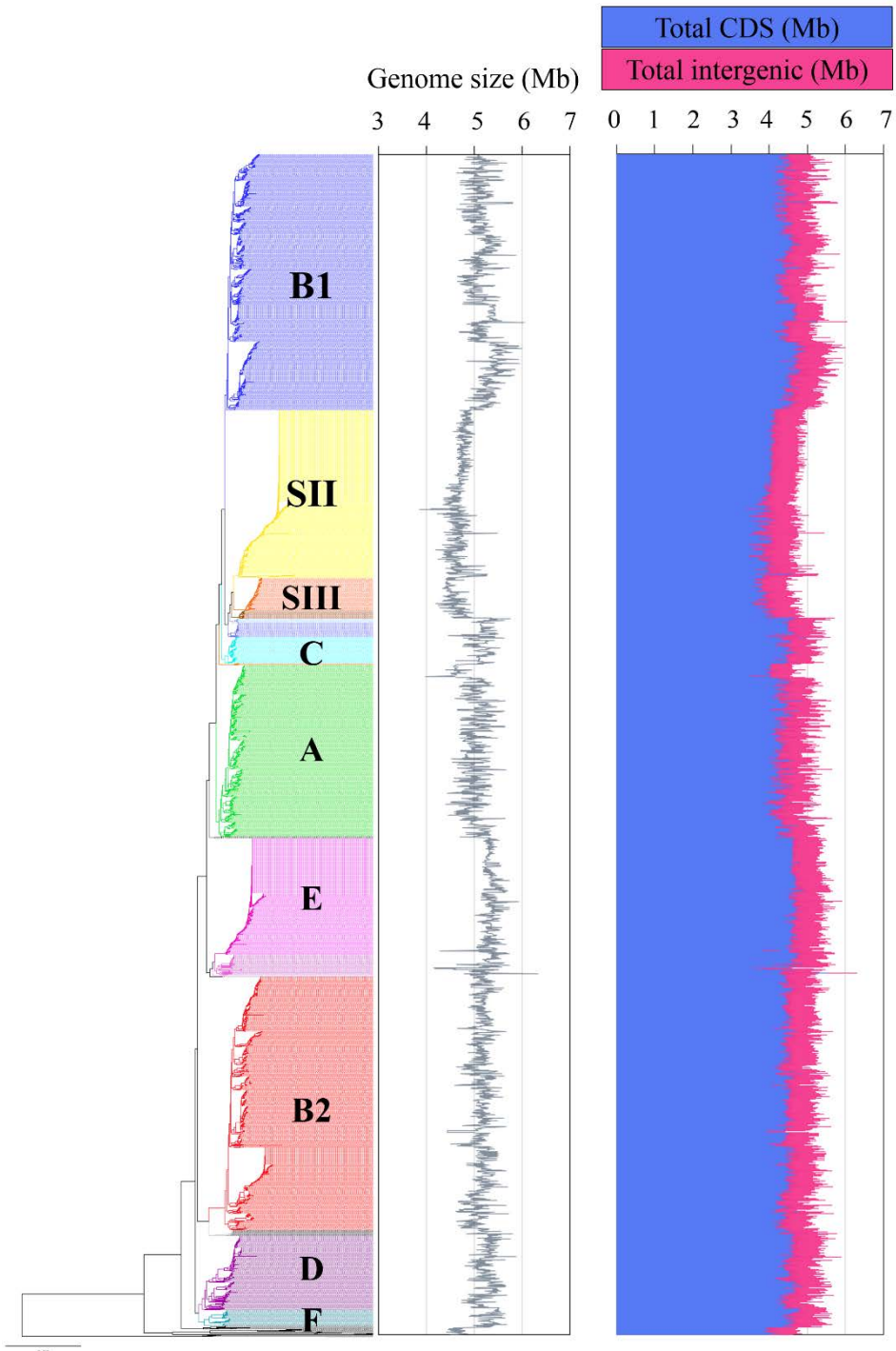

Figure 4. Variations of *E. coli* genome size and the proportions of intergenic sequences along the phylogenetic tree of the strains.

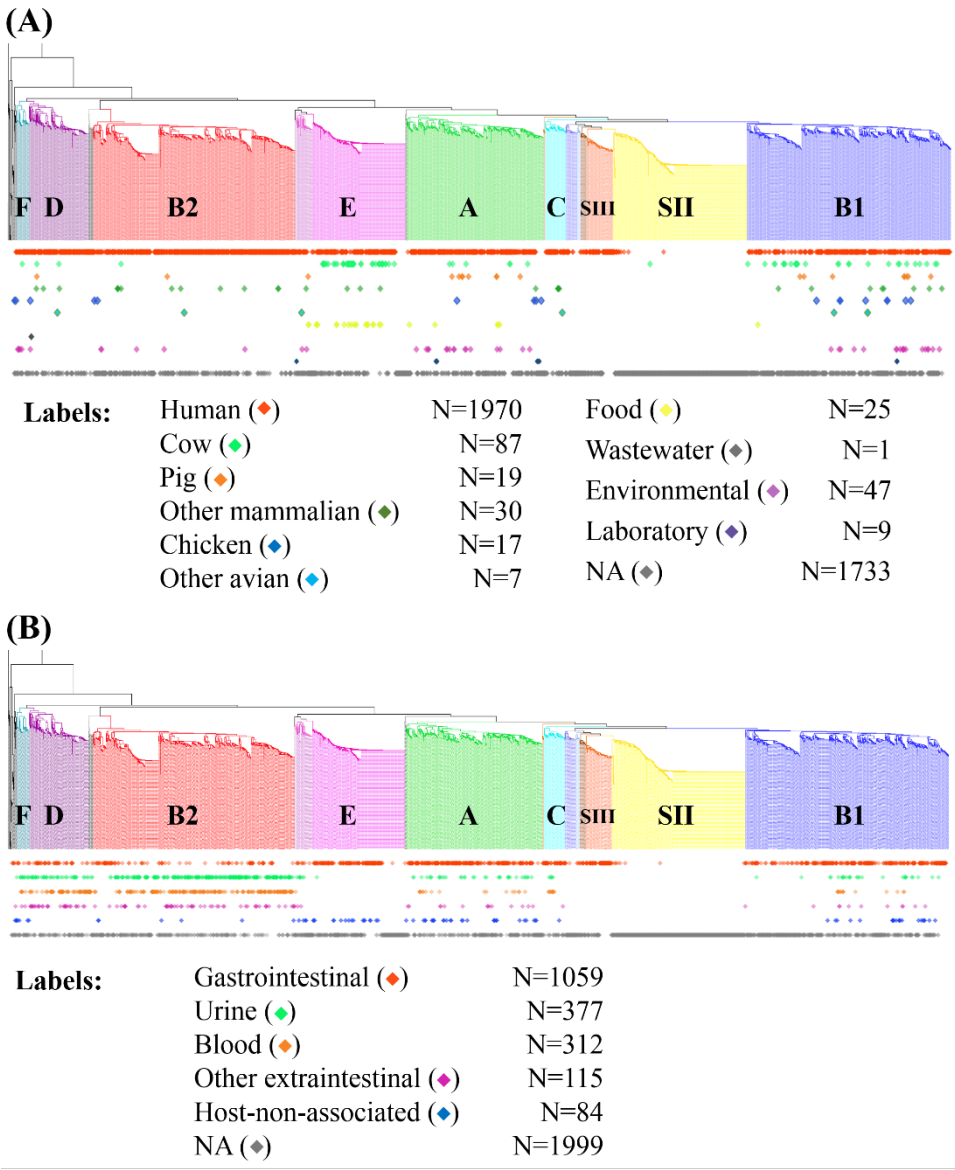

Figure 5. Distribution of isolation sources data in the strains used in this study. Host sources data (A) and Body site sources data (B) were plotted in the context of phylogenetic relationship of the strains.

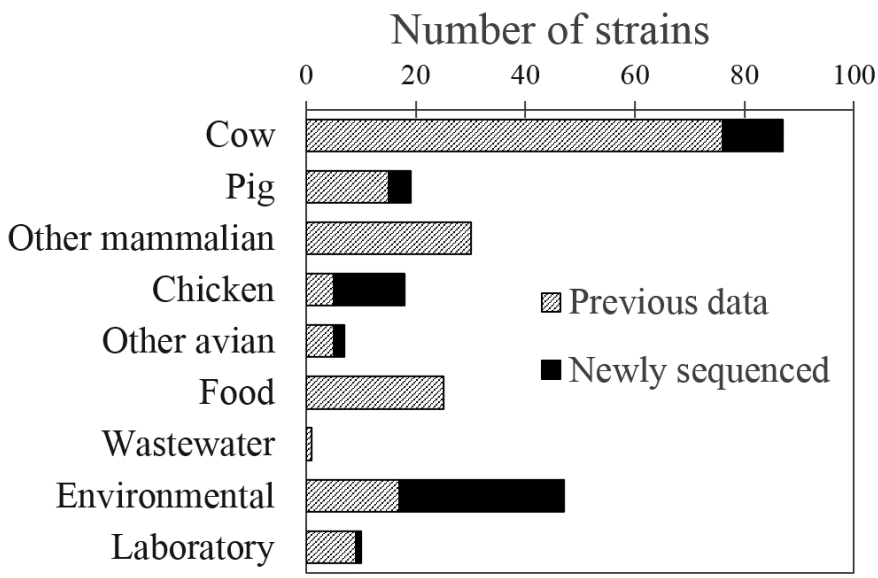

Figure 6. Contributions made by the genome sequences produced in this study to the availability of genome data of non-human isolates. Note that for environmental isolates and chicken isolates, the genome data deposited by this study comprised significant increment of available data.

2.3.2. Open pan-genome of *E. coli*

Clustering of orthologous protein-coding genes contained in the genomes of 3,946 *Escherichia* strains resulted in 84,468 clusters of orthologous genes. By counting the orthologous genes that were present in 3,909 *E. coli* and *Shigella* strains, 83,729 clusters of orthologous genes comprised the pan-genome of *E. coli*. Members of each cluster of orthologous genes could be regarded as intra-population allelic variants of the same genes, hence from here a cluster of orthologous genes was referred to just as a "gene". Characterization of the properties of *E. coli* pan-genome was performed using the 83,729 genes and 3,909 strains.

The size and openness of pan-genome were determined by two approaches. First, a widely employed analytical approach that uses sampling-dependent growth curve of pan-genome size was carried out. In this approach the pan-genome size is modeled by the number of strains (genome sequences) included in the dataset. Sampling-dependent pan-genome growth curve generated by 500 replication of random sub-sampling was obtained as **Fig. 7**. The sampling curve showed that pan-genome growth did not stop until 3,909 strains were analyzed. The fraction of singleton genes remained constantly in the range of 33-38% along the number of strains. Pan-genome size counted by ignoring singleton genes also indicated that the sampling of *E. coli* pan-genome would not be completed by sampling more strains. Interestingly, if the genes that were present in at least 1% of the strains were counted, the size of the collection of such genes converged quickly as the number of strains increased and ended up at 15,396 genes. This way of counting corresponded to the leftmost 1/100 fraction of the 'U-shaped' distribution of gene-frequency (**Fig. 10**).

This implied that among the gene-frequency spectrum (0-100%) the genes belonging to the right-side 99% of the spectrum were not participating in or influenced by the openness of *E. coli* pan-genome. Using PanGP (Zhao, et al. 2014) the sampling-curve of *E. coli* pan-genome was fitted to the equation of power law. The empirical *E. coli* pan-genome growth curve was fitted extremely well by the function, with R^2 value 0.99996 (**Fig. 8**). The exponent of the function was 0.33. As this value was in the range between 0 and 1, the pan-genome of *E. coli* obtained in this study should be considered as an open pan-genome.

To determine if the phylogenetic groups within *E. coli* also have their own open pan-genomes, pan-genome growth curve was analyzed separately for each phylogenetic group. Based on visual inspection the growth curves group A and B1 appear to have the largest pan-genomes (**Fig. 10**). Possible reasons for the larger pan-genome observed in group A and B1 might be (i) that the two groups have higher ecological diversity and/or (ii) that genome-sequenced strains were sampled from diverse lineages in the group, rather than focused on some specific lineage such as an important epidemic clone. In the intermediate range, the groups B2, C, D and F exhibited similar pan-genome size. The other 3 groups which are comprised exclusively by pathogenic strains, namely group E, SII and SIII, displayed markedly smaller pan-genome sizes. Explanations for observed small pan-genomes of these groups might be made based on (i) that these groups have very restricted ecological niches and/or (ii) that the genome-sequenced strains of these groups are concentrated to epidemic clones. The pan-genome growth curves of phylogenetic groups analyzed by power-law function to provide quantitative comparison of the shapes of growth curves. As shown in **Fig. 9**, pan-genome growth curves for all groups were able to

be closely approximated by power-law function. **Table 4** described the power-law functions derived from the pan-genomes of the phylogenetic groups. Counter-intuitively the results showed that the groups composed of very closely related strains such as group E, SII and SIII displayed larger *B* (the exponent of the function) values, when compared to the group A, B1, and B2 that contain greater strain diversity. The exponent of the function determines the rate at which the growth curve is flattened. The initial slope (*A*) of the growth curves were larger in the groups of more diverse strains. Interpretation of smaller *A* value and larger *B* value obtained for group E, SII and SIII would be that in these groups (i) the genomes are generally very similar to each other in terms of gene contents (smaller *A*) and (ii) the variable gene contents contained in each strain do not overlap much with the other strains.

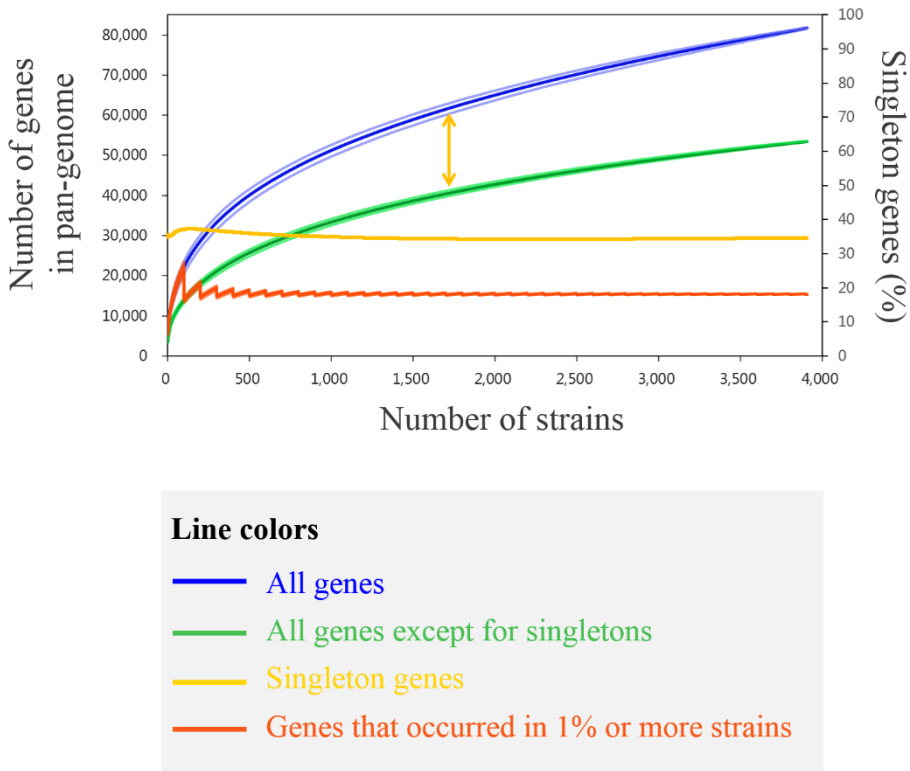

Figure 7. Sampling-dependent pan-genome growth curve of *E. coli/Shigella* strains. Axis on the left side was used for the number of all genes (blue), the number of non-singleton genes, and the number of genes with gene frequency value 1% or larger. The secondary axis on the right side was used for the proportion (%) of singleton genes.

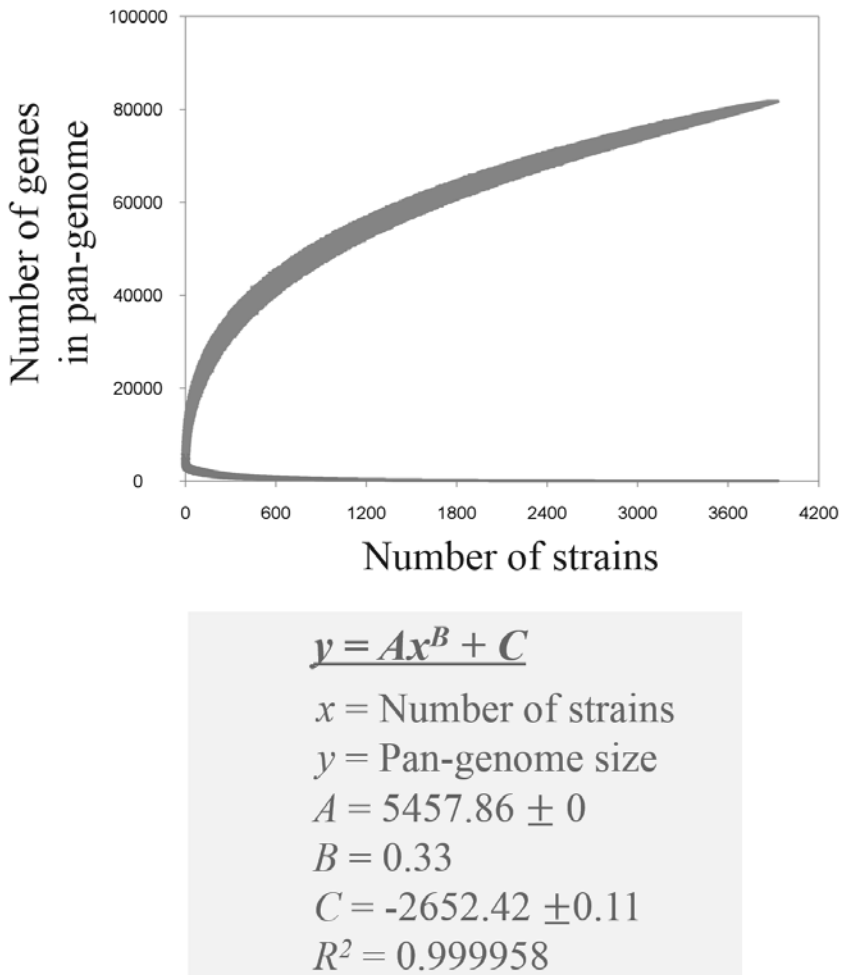

Figure 8. Power-law function fitted to the pan-genome growth curve of *E. coli/Shigella* strains. The statistical fitting and the figure was generated by PanGP.

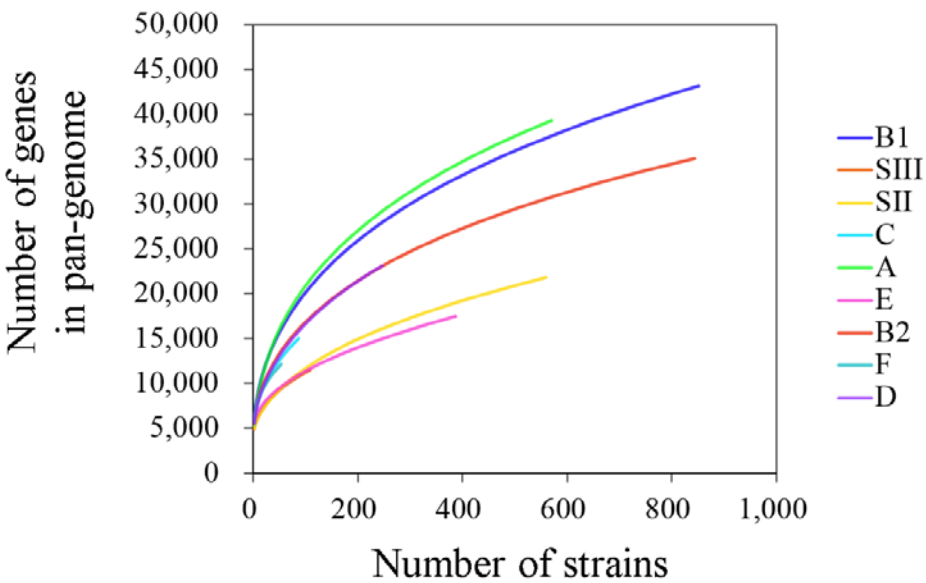

Figure 9. Pan-genome growth curves estimated separately for each phylogenetic group. The curves of phylogenetic groups ended at different horizontal value because the numbers of genome sequences analyzed for each phylogenetic group were different.

Table 4. Power law functions obtained for the pan-genome growth curves of phylogenetic groups of *E. coli.*

Group	*A*	*B*	*C*	R^2 of fitting
E. coli	5457.9	0.33	-2652.4	0.999958
A	4179.4	0.35	-866.1	0.999786
B1	3979.6	0.35	-104.2	0.999867
B2	3113.5	0.35	873.9	0.999798
C	2123.5	0.39	2658.4	0.999922
D	1896.4	0.43	2805.4	0.999872
E	806.6	0.46	4408.6	0.999914
F	3169.0	0.29	1866.4	0.999572
SII (*S. sonnei*)	1071.3	0.45	3139.8	0.999672
SIII (*S. flexneri*)	1119.5	0.42	3339.6	0.999739

The power-law function used in the fitting was: $y = Ax^B + C$. Values estimated for the parameter *A*, *B* and *C* and the R^2 of the function made up the columns in the table.

2.3.3. Statistical analysis of pan-genome gene frequency distribution

Each gene in the pan-genome had the number of occurrence between 1 and 3,909, or, population gene frequency between 1/3,909 and 1. Gene frequency distribution of the pan-genome of *E. coli* and *Shigella* was obtained as **Fig. 10**. It should be noted that the count for un-sampled genes (corresponding to the occurrence value 0 or gene frequency value 0) remained unknown in this analyses. Along the gene frequency spectrum, the most prevalent case was “singleton” genes that were present only in a single strain. ORFan genes made up 33.6% of the pan-genome (27,453 genes). Phyletically rare genes, defined as the genes that were present in 10 or less strains, made up 68.2% of the pan-genome (55,691 genes). Gene frequency distribution was also analyzed separately for *Shigella* strains, and for each phylogenetic group in *E. coli*. In the gene frequency distribution of *Shigella* strains two distinct features were recognizable (**Fig. 10**). First, *Shigella* had smaller fraction of genes in the left side of the gene frequency spectrum. Second, *Shigella* strains had a noticeable peak at the number of strains 551-565. Presence of this peak corresponded to the presence of 562 strains that all belonged to a recently emerged *S. sonnei* clone. When gene frequency distributions of different phylogenetic groups were compared, the groups displayed generally similar shapes (**Fig. 11**).

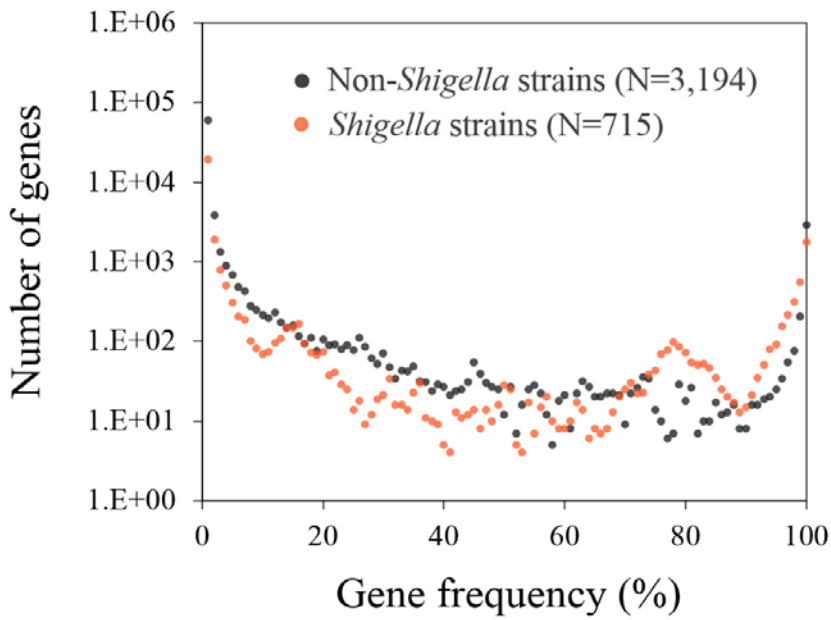

Figure 10. Gene frequency distribution of the pan-genomes of *E. coli* and *Shigella.*

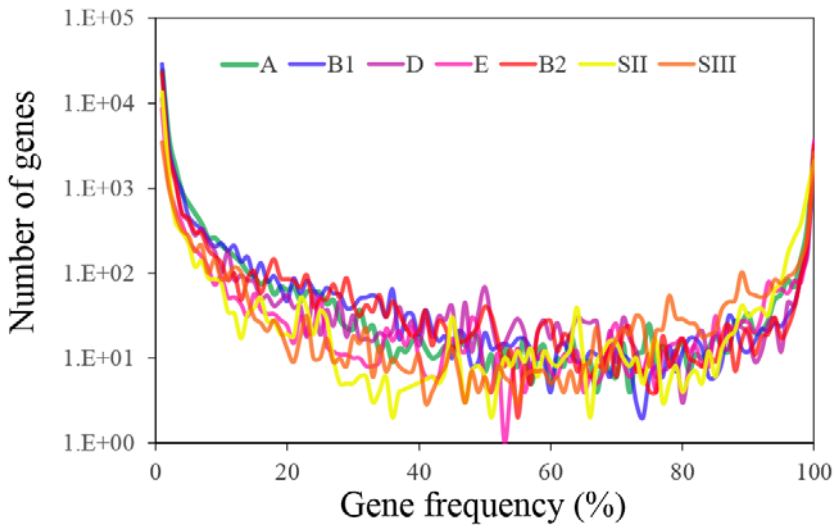

Figure 11. Gene frequency distribution estimated for each phylogenetic group.

The left-most part of gene-frequency distribution obtained from *E.* coli pan-genome could be fitted well by a power-law function $y = 24834\ x^{-1.376}$ with R^2 value 0.9973 (**Fig. 12**). Explicit interpretation of the parameters of this mathematical function to evolutionary genomic property of the population could not be made. The group of strains that have more genes unique to single strain and less genes shared in intermediate frequency could be expected to have lower exponent value. The exponents of power law functions fitted to the gene frequency distribution of phylogenetic groups of *E. coli* were compared with that of *E. coli* species (**Table 5**). Lower values were observed for the groups of strains that have restricted genetic diversity, such as group E, SIII (closely related *S. flexneri* strains) and SII (recently emerged *S. sonnei* clone).

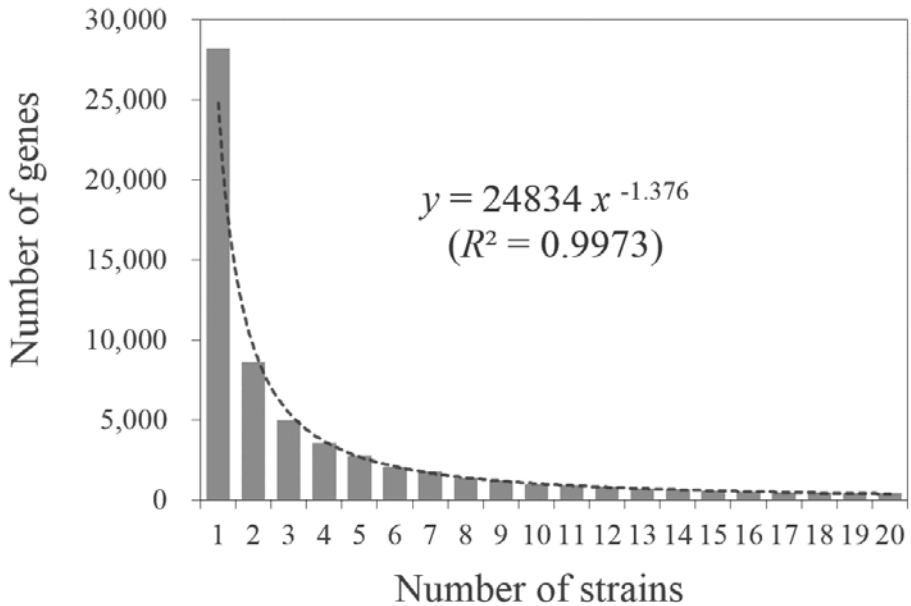

Figure 12. Statistical fitting of the left end of the gene frequency distribution of *E. coli* pan-genome. The power-law function used here was: $y = Ax^B$.

Table 5. Comparison between the statistical fittings results from the gene frequency distributions of different *E. coli* phylogenetic groups.

Group	Exponent of the power function (*B*)	R^2
E. coli	-1.376	0.997
F	-1.403	0.952
A	-1.423	0.995
B2	-1.425	0.998
B1	-1.427	0.994
C	-1.488	0.980
SIII (*S. flexneri*)	-1.536	0.980
E	-1.674	0.986
D	-1.692	0.995
SII (*S. sonnei*)	-1.775	0.985

From the power law functions fitted to the left-ends of gene frequency distributions of each phylogenetic group, the exponential component *B* in the function $y = Ax^B$ was compared in the table. The table was sorted by descending order according to the value of the value of *B.*

2.3.4. Evolutionary rate of pan-genome growth

An alternative version of pan-genome growth curve where the horizontal axis represent the accumulated core-genome phylogenetic diversity of the given strains, instead of the number of the strains, was designed and applied to the pan-genome of *E. coli*. Before this analysis, size and openness of *E. coli* pan-genome have been addressed by analysis of gene frequency distribution and pan-genome growth curve. Both gene frequency distribution model and pan-genome growth curve model used the number of strains as a key parameter that can affect the pan-genome size. One disadvantage of these models was that depending on the sampling strategy the same number of strains can be composed of strains with different levels of diversity. For example, the number of strains would be the same for 100 strains that belong to a specific clone and for 100 strains that are dispersed evenly throughout multiple clades. As a result, the characterizations obtained based on such approaches should be interpreted as the property of the pan-genome of specific dataset (a list of genome-sequenced strains) and should not be taken as a property of the pan-genome of the species (an evolving population). In the method designed in this study, the number of strains was replaced by the accumulated core-genome phylogenetic diversity. The accumulated core-genome phylogenetic diversity was measured by the sum of branch lengths that were needed to connect all strains in the phylogenetic tree. The phylogenetic tree should be derived from core-genome alignment. The pan-genome of *E. coli* and the pan-genomes of each phylogenetic group were analyzed in this way.

As shown in **Fig. 13**, the relationship between pan-genome size and the core-genome diversity of the subsampled strains appeared to be linear in *E. coli* genome

data. Linear function f(x) = $ax + b$ was fitted to the observed data points. The slope of the fitted linear regression corresponds to the relative increment of pan-genome size for unit increment of phylogenetic diversity measured in terms of branch length (substitution per site) of core-genome phylogenetic tree. Branch length was defined as number of substitutions per site and pan-genome size was defined as the number of genes. Thus the slope of function represents the relative rate of gaining new genes versus gaining new substitution in the core-genome, for an evolving species. Given that linear relationship between the core-genome diversity and pan-genome size, the slope value derived from diverse phylogenetic groups and species were compared. The linear plots obtained for the pan-genomes of phylogenetic groups within *E. coli* were visualized together in the **Fig. 14**. The end-point of each line is the end-point values of core-genome diversity and pan-genome size. Groups of strains that occupy smaller total phylogenetic diversity (end-point x value) generally showed more rapid slope in the plot (**Fig. 15**). This association between the overall phylogenetic diversity of the group and the rate of pan-genome growth per unit increase of phylogenetic diversity was not a systematically generated bias. When the pan-genome of randomly defined group of strains was analyzed (the grey circles in the Fig. 15), the negative association was much weaker than what was observed for phylogenetic groups. At least three possible explanations occurred for this trend. First, there could be a systematic bias in our approach. Second, the group C, E, F, SII and SIII are actually more rapidly expanding their genetic repertoire than the other groups. Finally, this could also be explained by the assumption that *E. coli* genome evolution in short time-scale (among very closely related strains) is dominated by acquisition of new genes, while in long time-scale (among wide variety of strains) the impact of orthologous sequence diversification catch up the impact of gene content diversification.

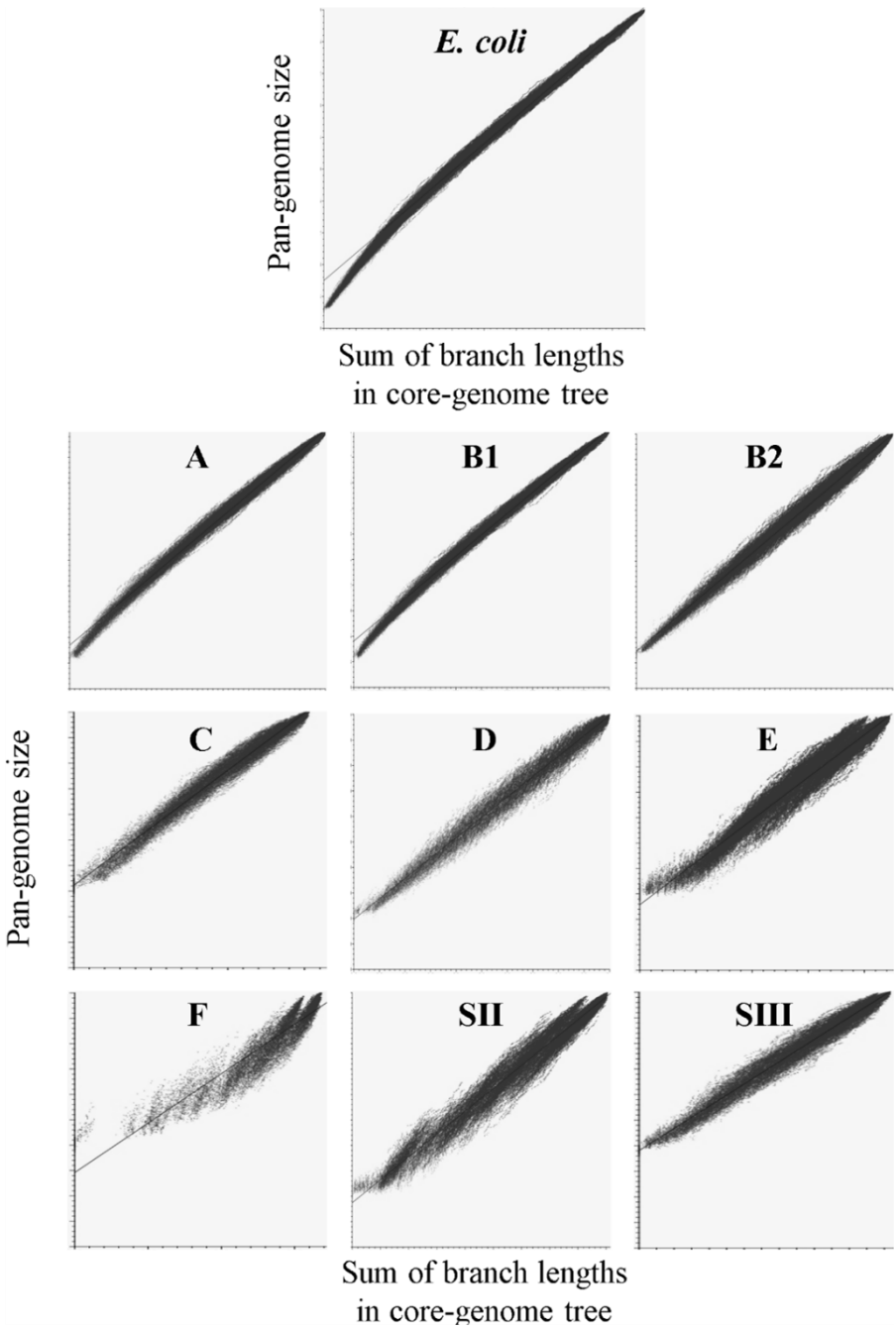

Figure 13. Relationship between pan-genome size and the phylogenetic diversity of the strains.

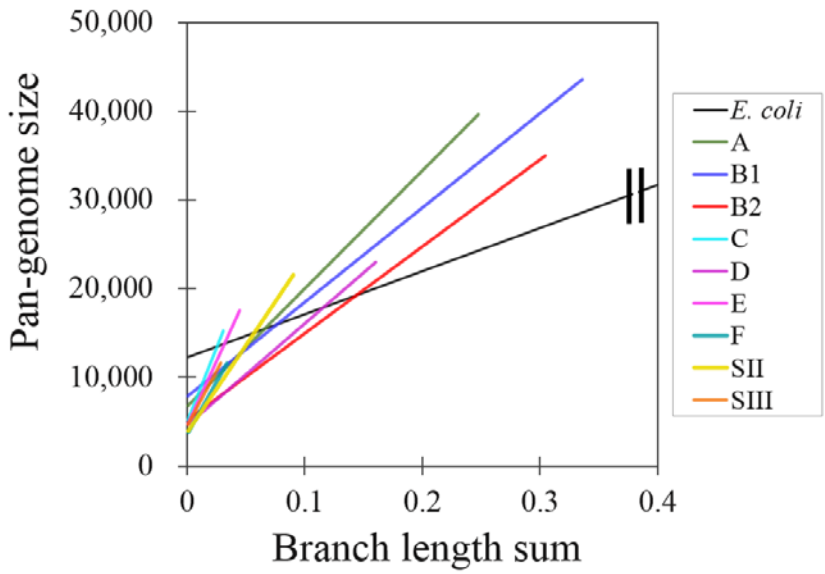

Figure 14. Comparison of the linear regressions obtained for the pan-genomes of different groups. The plot for *E. coli* was cut because the endpoint *x* value of *E. coli* was much larger than that of any individual phylogenetic groups.

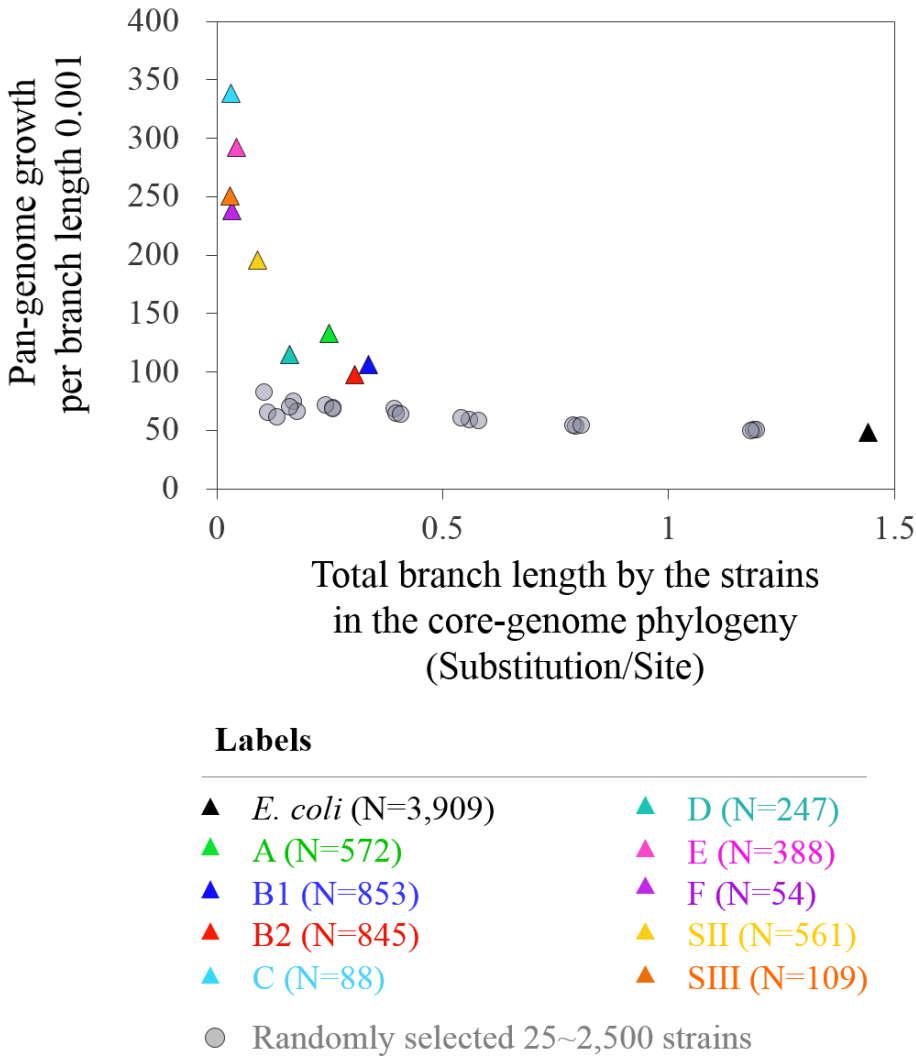

Figure 15. Negative association between the slopes of pan-genome growth per phylogenetic diversity and the total phylogenetic diversity of the group of strains. Data points obtained from randomly selected subset of *E. coli* strains did not show the same trend (grey circles) with the phylogenetic groups (colored triangles). Notably, group B2 and group D deviated slightly downward from the general trend.

2.3.5. Phylogenetic and population genetic structure inferred from genome data

Evolutionary relationship between the strains were determined by phylogenetic analysis of core-genome sequences. The result offered a more comprehensive phylogeny compared to the previously reported phylogenies of *E. coli* strains that were based on a few number of genes (Clermont, et al. 2013; Turrientes, et al. 2014) and the genome-scale phylogenies that included a small number of strains (Touchon, et al. 2009; Zhang and Lin 2012). As shown in **Fig. 16** the phylogenetic tree reconfirmed the presence of phylogenetic structure within the species. Phylogenetic groups were assigned to the strains were based on the conventional assignments of the strains that were covered in the previous literatures. In addition to the previously known phylogenetic groups, 6 miscellaneous groups were added systematically to cover all 3,946 *Escherichia* strains. Assignment of the group followed the rule that all groups should be monophyletic. *Shigella* strains were clustered into 4 clades except for *S. dysenteriae* strains that appeared among non-O157 STEC strains near the root of group E. Non-*Shigella E. coli* strains were divided into 7 major groups plus 6 miscellaneous groups newly defined in this study. The 6 miscellaneous groups are: two monophyletic groups of strains that diverged just before the divergence of SII, SIII, SIV groups ("Near-*Shigella*-B1 I" and "Near-*Shigella*-B1 II"); a clade that consisted of non-O157 STEC strains and *S. dysenteriae* strains ("STEC/Sd"); a clade that diverged just before the common ancestor of group A and B1 ("B1/A sister"); a clade that diverged from the ancestor of group E ("E sister"); a sister clade of group B2 ("B2 sister"). Evolutionary relationship of the groups defined here was schematically represented in the **Fig. 17**.

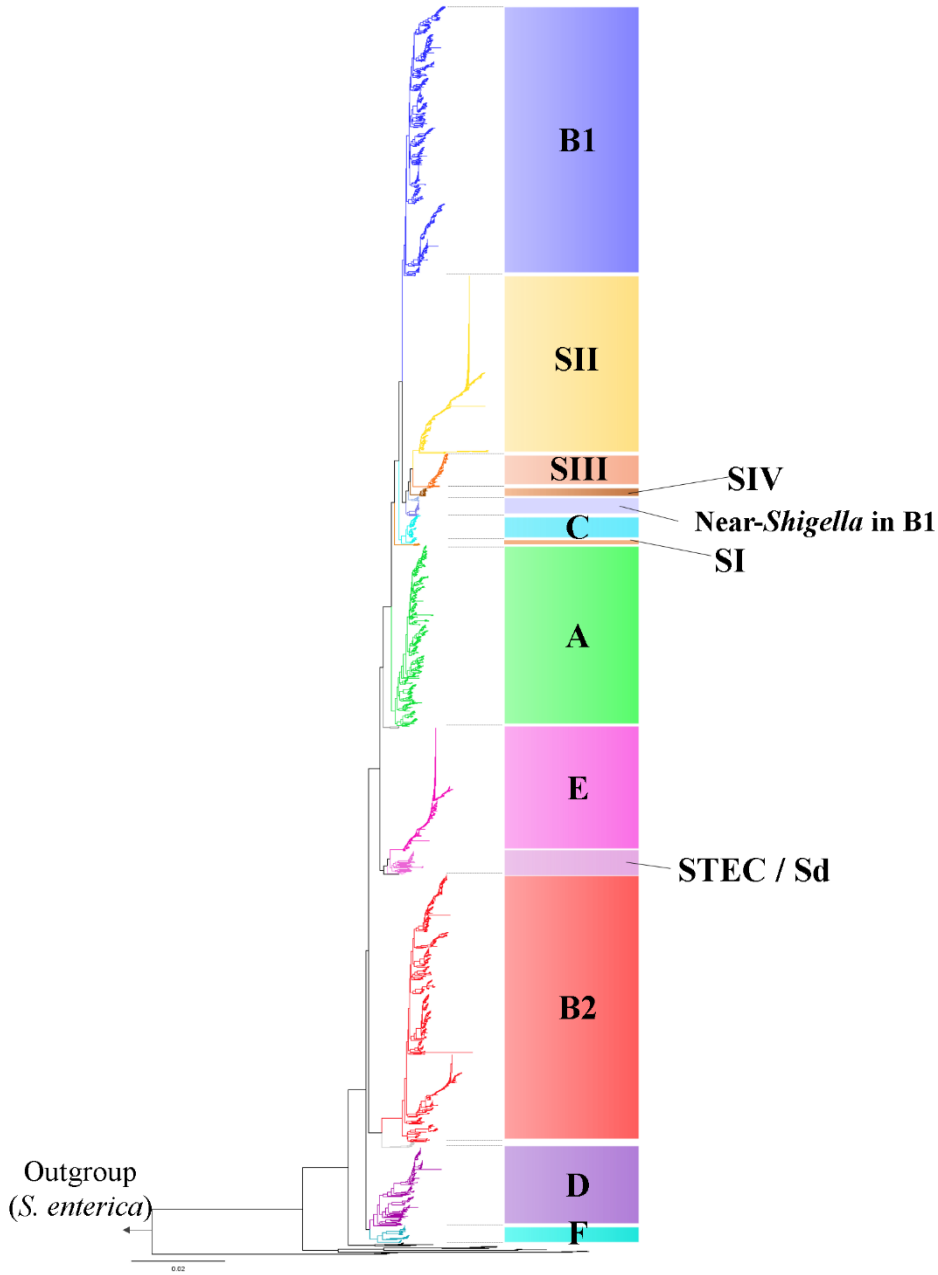

Figure 16. Maximum-likelihood phylogenetic tree of the strains in *Escherichia* genus. The tree was inferred for non-recombinant core genes. Strains were classified into the groups according to the monophyly of the group of strains based on in this tree.

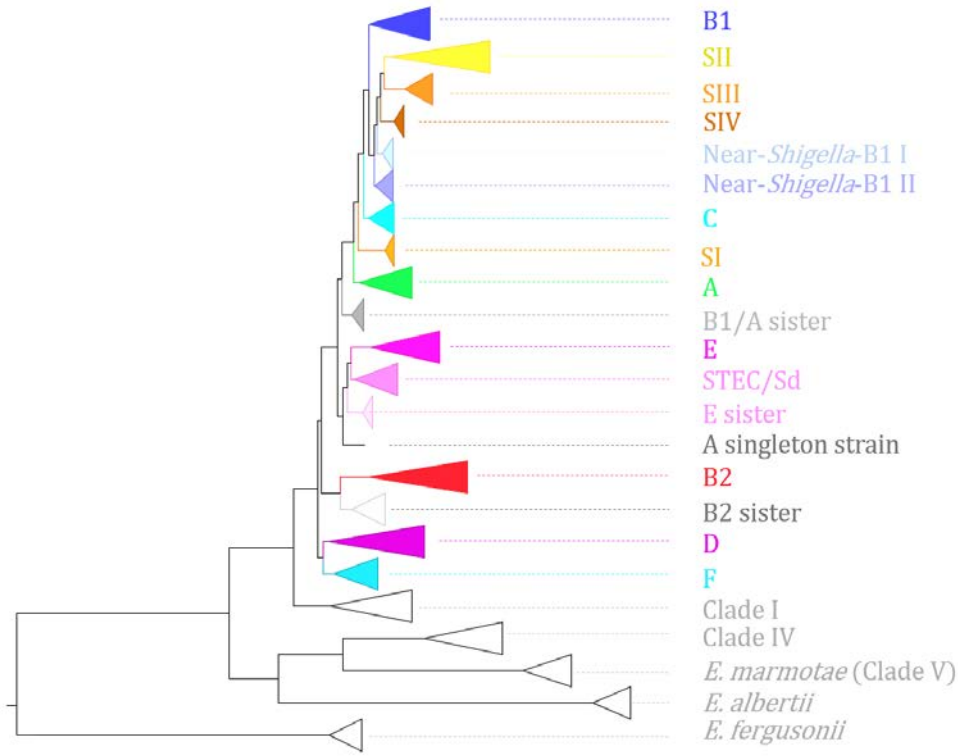

Figure 17. Schematic representation of the evolutionary relationships of the phylogenetic groups. The scheme contained 7 major *E. coli* groups, 4 *Shigella* groups, and 6 miscellaneous groups defined in this study, plus the species and clades other than *E. coli.*

Presence of subspecies level population structure was tested by using population genetics method to supplement the results of phylogenetic analysis. Alleles state at the bi-allelic SNP loci in the genomic regions conserved in all strains were used to infer population genetic structure of the strains. Whether or not the population structure exists among the strains and if the population structure matches with the phylogenetic clustering of the strains were the questions behind this analysis. A non-parametric method for subpopulation clustering, discriminant analysis of principal components (DAPC) was used to cluster 1,897 genomes into the given number of subpopulations. Since the true number of subpopulations was unknown, analysis was repeated for the assumption of the number of subpopulations (K) from 2 to 11. Mixed individuals did not appear at when $K < 8$. When assumption of K was 8 or larger, individuals with mixed affiliation appeared (**Fig. 18**). Clustering of the strains into 2-7 subpopulations generally resulted in the splits between the strain that were congruent with the phylogenetic clustering as seen in the **Fig. 18**. An exceptional incongruence was observed when K was set to be 7, as the two of the subpopulations were defined in the internal branch of group B1 and B2, respectively. When assumption of the number of subpopulation was greater than 7, DAPC resulted in the presence of mixed individuals. The most realistic number of subpopulations cannot be confirmed in this analysis. Tests of genetic isolation should be performed to estimate to which degree the detected phylogenetic clusters or subpopulations are evolving independently from each other.

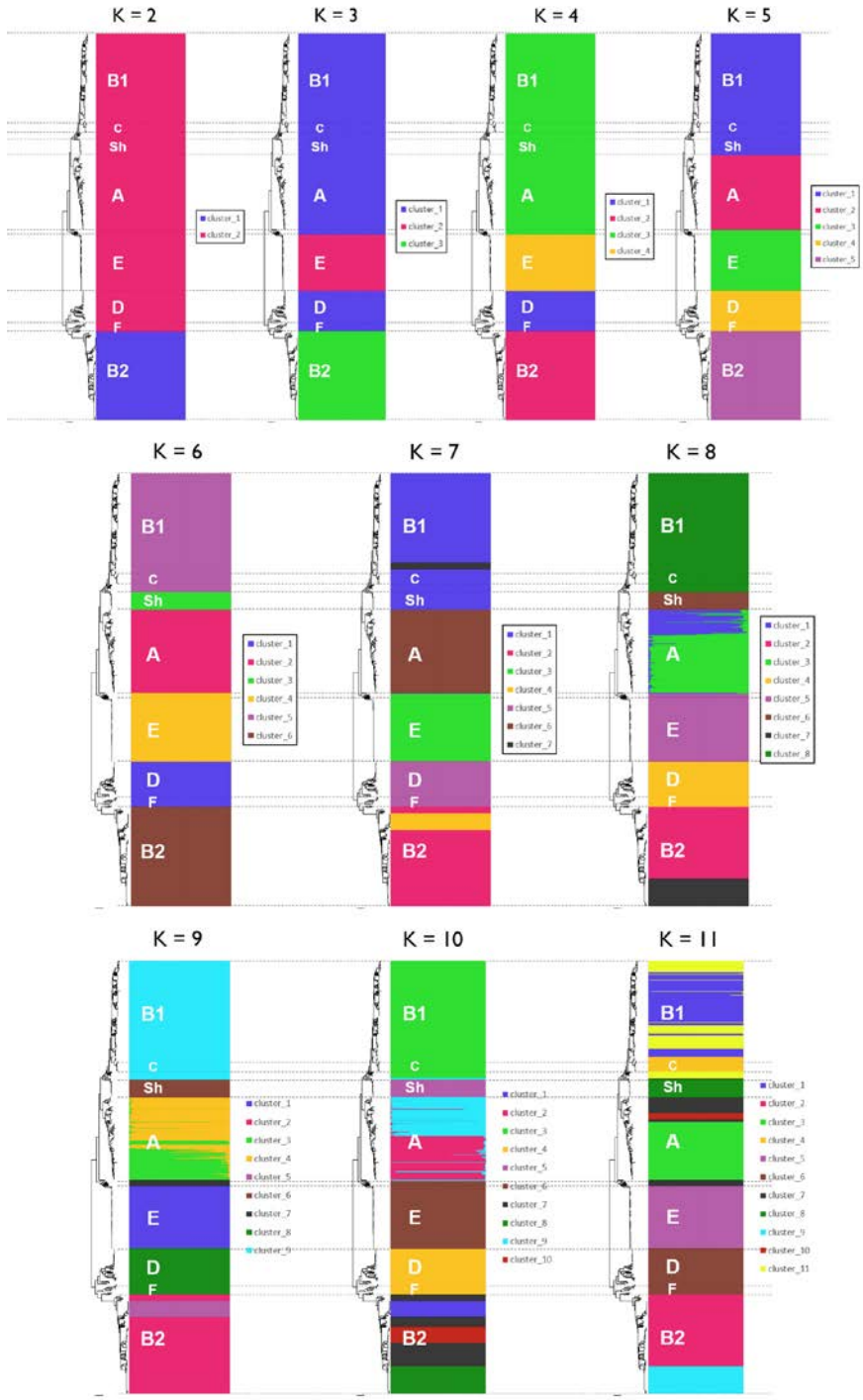

Figure 18. Subpopulation clustering and posterior probabilities of affiliation made for each strains by DAPC method.

Existence of subspecies level structure was detected both by phylogenetic clustering and population genetic clustering. The nature of the within-species groups that were recovered was still unclear. One clue to that was obtained from genome-wide linkage analysis. Genome-wide linkage analysis results was mainly discussed in the subchapter 2.3.7. Decay of LD depending on the physical distance between SNP sites was observed in the LD-decay plots shown in **Fig. 19**. The decaying of LD occurred due to the presence of recombination events that broke the linkage between the markers. When LD-decay plots for each phylogenetic group was compared with the LD-decay plot of the species *E. coli*, LD within individual phylogenetic groups was maintained at higher level than LD within the species. This implied that the recombination's activity in breaking LD was less frequent after the divergence of phylogenetic groups than before the divergence of phylogenetic groups. In other word, somewhat clonal nature of phylogenetic groups were implied by the LD decay analysis.

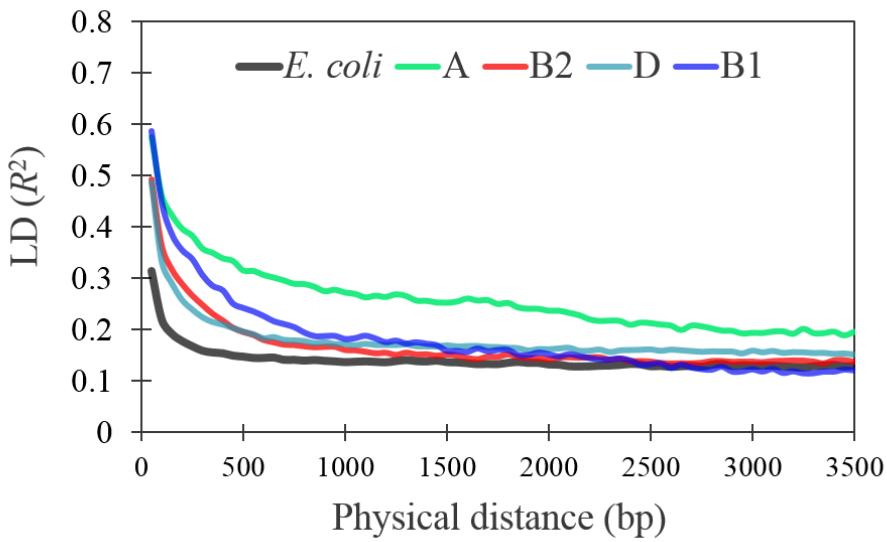

Figure 19. LD-decay plots estimated for *E. coli* and for each phylogenetic group within *E. coli*. The figure demonstrated that LD-decay within each phylogenetic group was slower than what was seen in whole species level LD calculation.

2.3.6. Intra-specific sequence diversity in the pan-genome of *E. coli*

One previous study of genetic diversity of *E. coli* reported that the core-genome of *E. coli* had nucleotide sequence diversity (as measured by average difference per site) of in the distribution that has a peak at 2% (Kaas, et al. 2012). In our analyses, overall sequence diversity distribution of core and non-core genes were compared. The question behind this analysis was whether or not core genes are more conserved than non-core, dispensable genes. For *E. coli* the genes in the population frequency categories of 99-100%, 97-99% and 75-97% all showed similar range of sequence diversity. Genes that were shared by the majority of the strains have narrow sequence diversity distribution. For rare genes, the distribution's peak appeared at lower value and at the same time the distribution was wider (**Fig. 20**). On average phyletically rare genes had smaller sequence diversity than core genes but there were some cases of phyletically rare genes that exhibited great sequence diversity that were not observed in the core genes. The same trend was observed in the pan-genome of *S. enterica* and *B. fragilis* (**Fig. 20**). Observed smaller average sequence diversity of non-core genes could be expected under neutral evolution, since the core-genes have been diversified in the population for longer time compared to the non-core genes. The wider distribution of non-core gene's sequence diversity might be explained by the exertion of diversifying positive selection to some of the dispensable genes, or by the incorporation of distant xenologous sequences by some strains (e.g. horizontally transferred genes). Determination between the two possibilities cannot be made in this analysis.

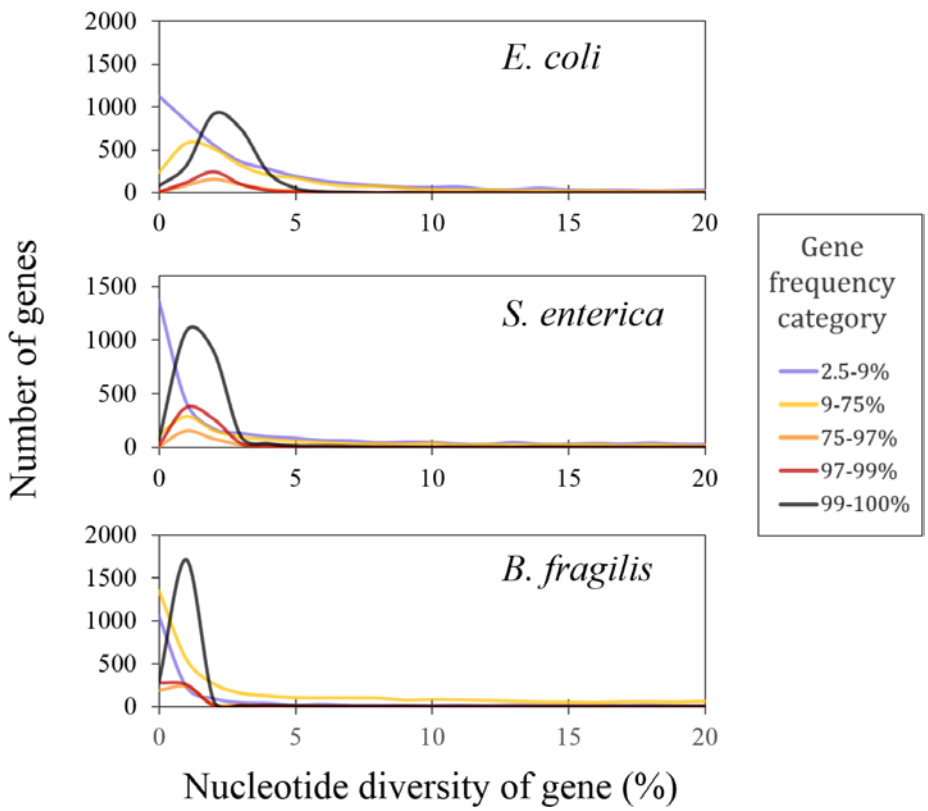

Figure 20. Distribution of nucleotide diversity of the genes in different gene frequency categories, from core genes to rare genes.

Sequence diversity spectrum of *E. coli* core genes spanned from 0% to 32.3%. Distribution of the nucleotide diversity of core genes was inspected by their chromosomal distribution (**Fig. 21**). Hyper-variable genes (>10%) were found in 5 separated genomic loci and the loci were distributed sparsely. In one locus, 3 hyper-variable genes, *yfc*P, *yfc*R and *yfc*S, were found together in proximity. The operon *yfcOPQRSTUV* is known to be unexpressed under normal culture conditions and the expression of the operon is known to promote the adhesion of *E. coli* to eukaryotic epithelial cells (Korea, et al. 2010). It is very interesting to observe that the greatest sequence diversity among *E. coli* core genes was observed in the operon that promote the adhesion of this bacteria to the Eukaryotic host. Observed diversity of body sites and host range of *E. coli* might be related with the sequence diversity of this operon. On the other extreme, there were 17 hyper-conserved core-genes that showed absolutely no sequence diversity. Other than ribosomal subunit protein genes, a murein lipoprotein of peptidoglycan cell wall, a translation initiation factor, a global translational regulator gene (*csrA*) (Dubey, et al. 2003), and a membrane-bound ATP synthase were included in the hyper-conserved genes.

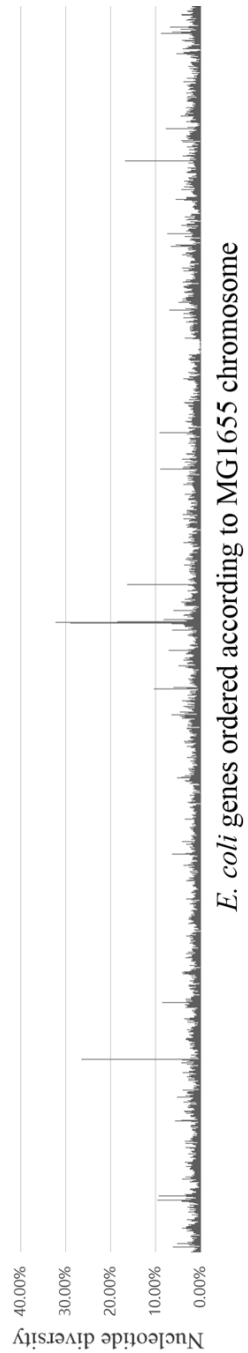

Figure 21. Distribution of nucleotide diversity on the chromosome of MG1655 strain. Hyper-variable genes appeared as peaks.

2.3.7. Analysis of gene content variation

The pan-genome gene presence/absence matrix was visualized as a heat-map to provide an overall look of the gene content diversity of *E. coli* strains (**Fig. 22**). The genes that were present in all strains or in a single strain were excluded from the heat-map. Phylogenetic tree was aligned to the heat-map in the left side to see the presence of the genes that were associated with phylogenetic clade. Presence of group-specific genes was indicated in the **Fig. 22**. To systematically extract the group-specific genes, the genes that satisfied intra-group gene frequency cutoff 0.9 and the out-group gene frequency cutoff 0.1 were selected. Except for group A, for all major groups one or more genes specific to the groups were discovered (**Fig. 23**). By comparing the gene contents of *E. coli* and the neighboring species using the same criterion, the genes that are signature of *E. coli* were detected too. Interestingly, the number of genes that were signature of group SII, SIII or E strains were larger than the number of genes specifically present in the *E. coli*. These group-specific or species-specific genes are potentially useful for diagnostic purpose (Clermont, et al. 2013). Moreover these specific genes are the candidates of the genes that promoted their ecological divergence.

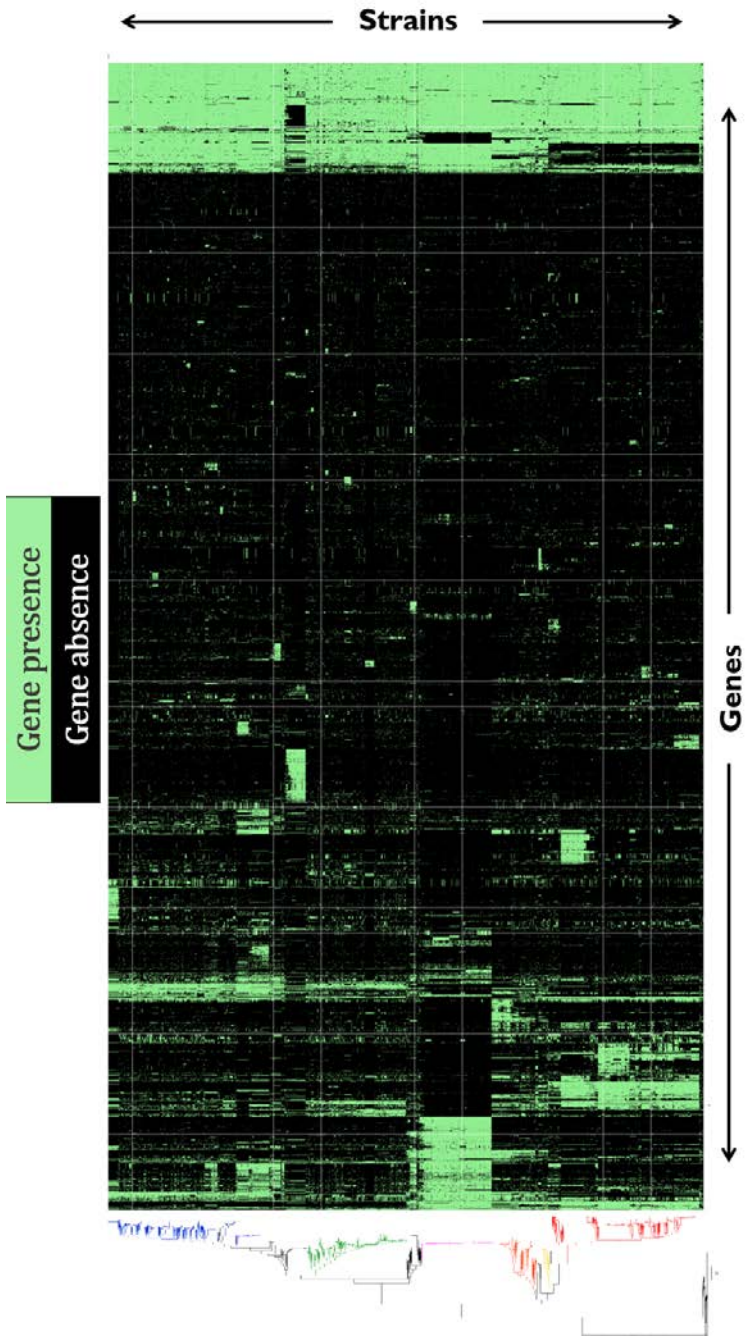

Figure 22. Presence/absence distribution of dispensable genes among *E. coli* strains. The strains were sorted according to the phylogenetic tree shown in the bottom side.

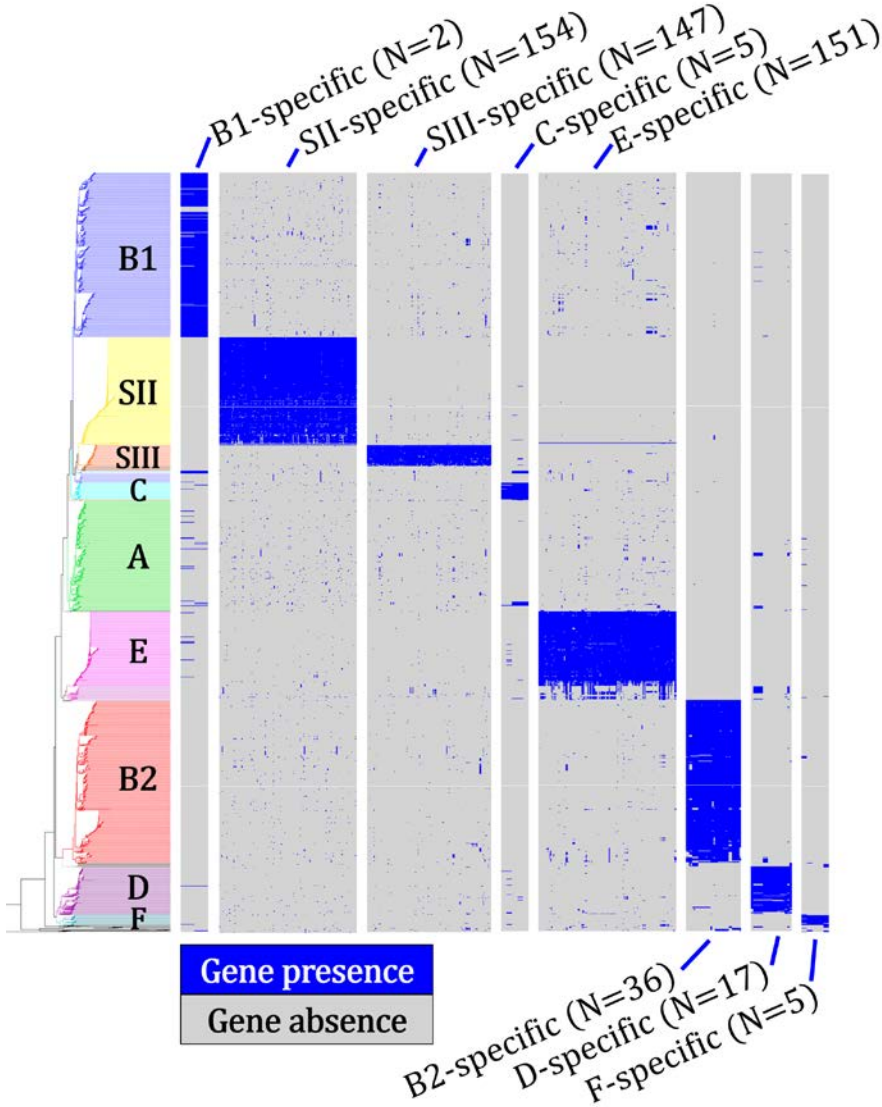

Figure 23. Group-specific genes in *E. coli* pan-genome. Group SII, SIII and E had the largest number of genes diagnostic to them. In contrast, group B1 only had 2 genes diagnostic to them and group A did not have any diagnostic gene.

Relationship between the gene content divergence and sequence divergence was another question that could be answered by pan-genome data. Presence of correlation between sequence distance and gene content difference were tested by exhaustive pairwise measurements of gene content dissimilarity and sequence dissimilarity for all possible pair of strains among 3,909 *E. coli*/*Shigella* strains. Gene content dissimilarities were measured by Bray-Curtis distance formula. Sequence dissimilarities were calculated as a simple per-site difference of core-genome sequences. Intra-species distribution of gene contents dissimilarities had a unimodal distribution (**Fig. 24**). The peak was found at 0.23, meaning that if you peak any two *E. coli* genomes and estimate the Bray-Curtis gene content dissimilarity the most frequent value to be obtained would be 0.23. The correlation between the two values were positive as shown in the **Fig. 25**. Frequently, pair of strains that were not diverged in terms of sequence difference showed significant gene content dissimilarity. The opposite case, pair of strains that did not show gene content dissimilarity and significant sequence divergence, was not observed.

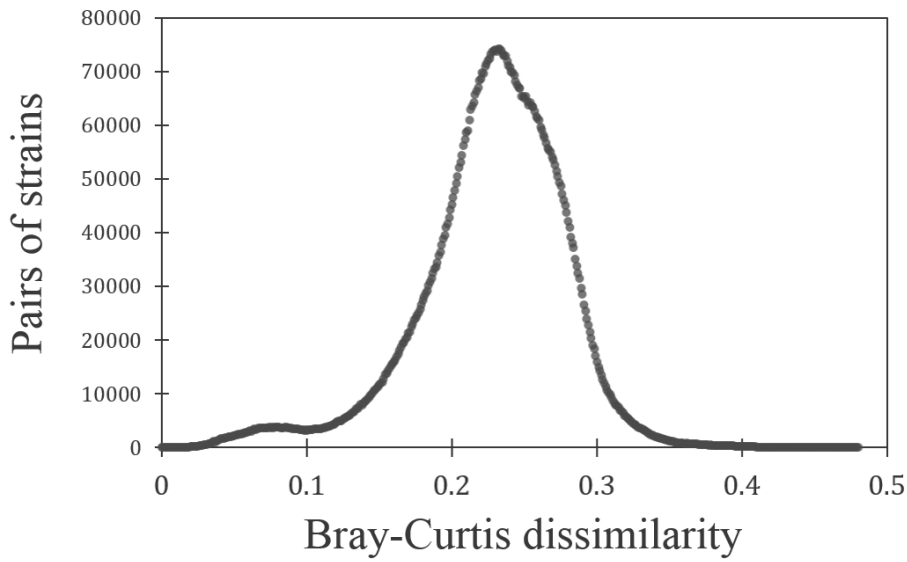

Figure 24. Distribution gene content dissimilarities within *E. coli*. Gene content dissimilarity between pair of genomes were quantified by Bray-Curtis index of dissimilarity. The central peak was found at 0.23.

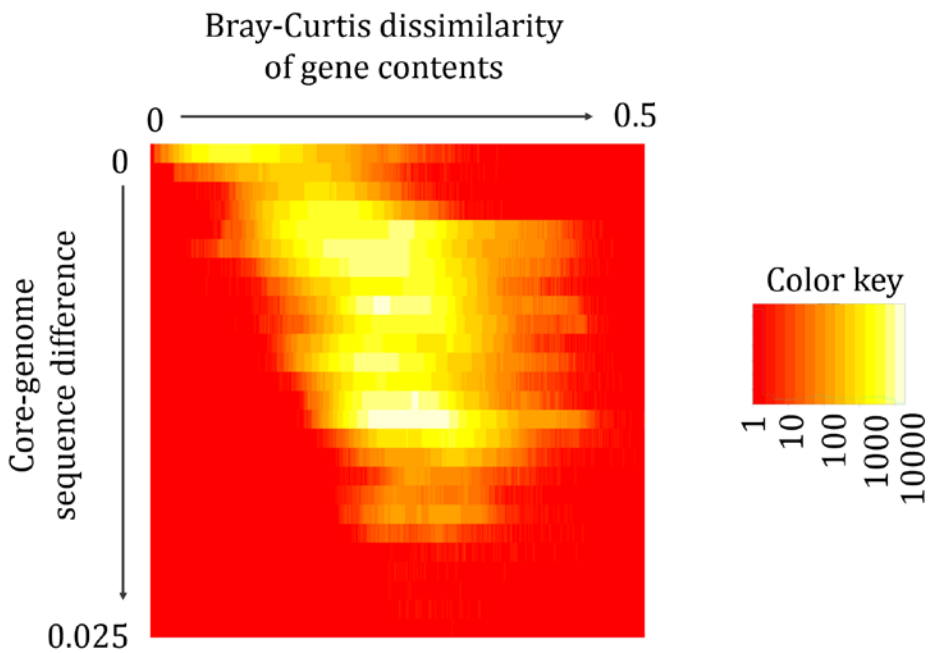

Figure 25. Relationship between the gene content dissimilarity and core-genome sequence difference. The heat-map is based on the counts of the pairs of strains that exhibited certain combination of core-genome sequence dissimilarity and Bray-Curtis dissimilarity. Count values were obtained from all possible pairwise comparison between 3,909 genomes.

2.3.8. Conservation of synteny and linkage over long distance

Conservation of genome-wide synteny among *E. coli* genomes was reported in one previous study (Rasko, et al. 2008). However comparative evaluation of inter-group and inter-species conservation of gene synteny has not been reported for *E. coli*. In this study, conservation of synteny was evaluated for *E. coli* core genes across various phylogenetic distances. Using the chromosomal order of core genes in the MG1655 strain as a fixed reference, chromosomal core gene orders in the completely assembled strains belonging to various phylogenetic groups were plotted in **Fig. 26**. In the plots strains of *E. coli* were shown to have remarkably preserved core gene orders even in the presence of frequent gain and loss of dispensable genes. In contrast, when the gene order of MG1655 was compared with that of the strains belonging to *E. fergusonii*, *E. albertii* or the other clades in the *Escherichia*, the level of synteny preservation was apparently poor. Observation of stable and *E. coli*-specific core gene orders implied that although *E. coli* genomes are dynamic as demonstrated by open pan-genome, some sort of un-interrupted 'backbone' has persisted among *E. coli* genomes.

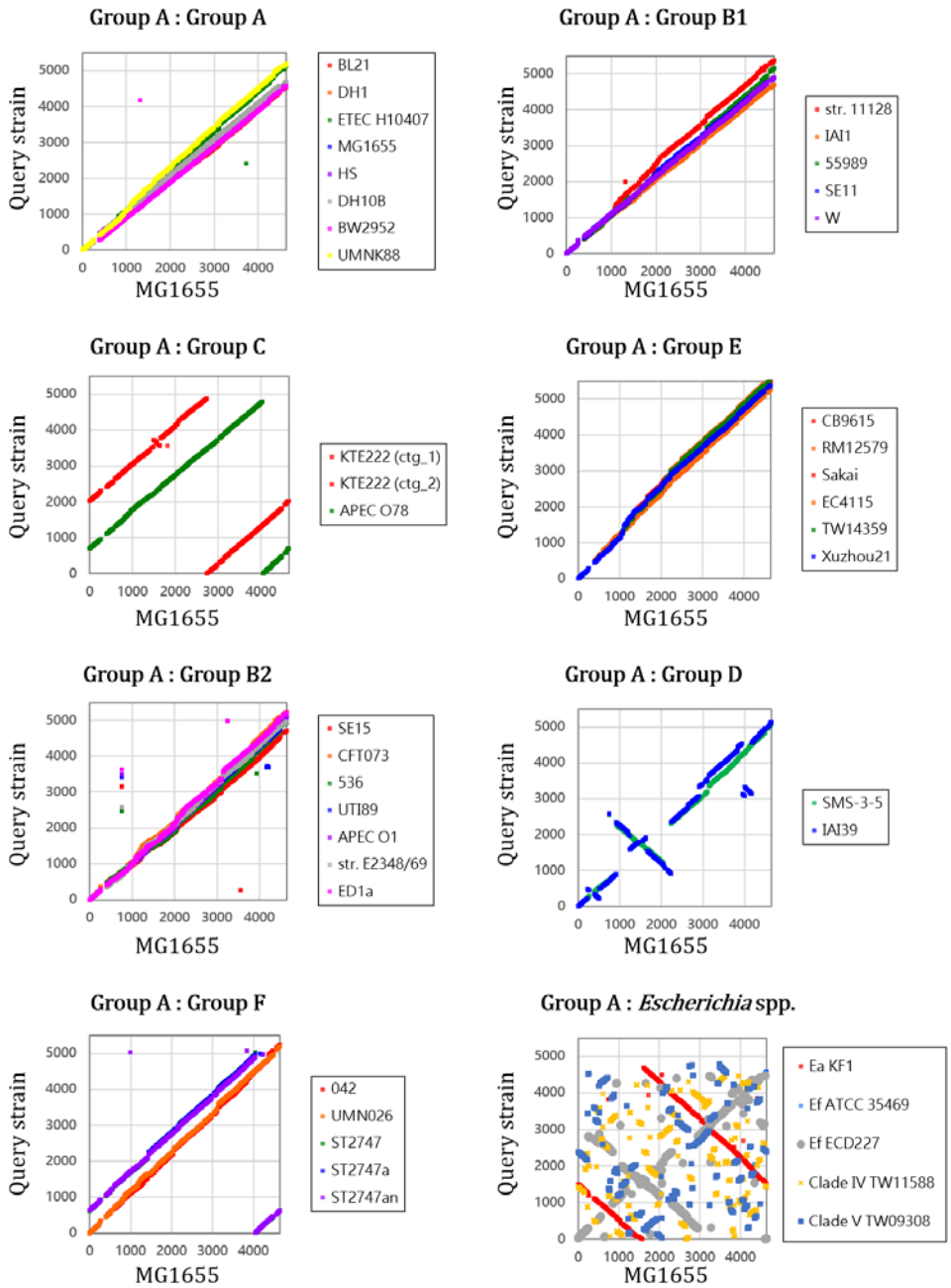

Figure 26. Conservation of synteny within *E. coli.* The genome of MG1655 was used consistently as a reference.

Concept of ‘clonal frame’ has been proposed as a unique phenomenon of bacterial genomes (Sarkar and Guttman 2004; Tibayrenc and Ayala 2012). The idea of clonal frame is a very long haplotype block spanning entire chromosome of bacteria. Such phenomenon is not possible in sexual organisms that undergo meiotic crossover. Unlike meiotic recombination, the impact of bacterial homologous recombination affects short stretches of chromosome and its impact does not accumulate over physical distance. Using the genome-wide core SNP data the linkage between SNP sites was analyzed to evaluate the clonal backbone hypothesis in *E. coli* genomes. Decay of LD between pair of SNP sites over the growth of inter-marker physical distance was examined. In the **Fig. 27** both *D*’ and R^2 statistics of LD showed constant decay over the short physical distance from 0 Kb to 10 Kb. In large scale, LD did not decay constantly when physical distance from 0 Kb to 800 Kb was examined. After 100 Kb, LD was maintained at high level as *D’* fluctuated in between 0.80 – 0.84. Quick decay of linkage within the 10 Kb distance demonstrated the impacts of frequent homologous recombination events in the history of *E. coli.* Maintenance of *D*’ value at certain level over longer distance demonstrated the presence of clonal frame phenomenon in *E. coli* genomes. Using a whole chromosome-wide linkage heat-map, the clonal frame of *E. coli* was visualized (**Fig. 28**).

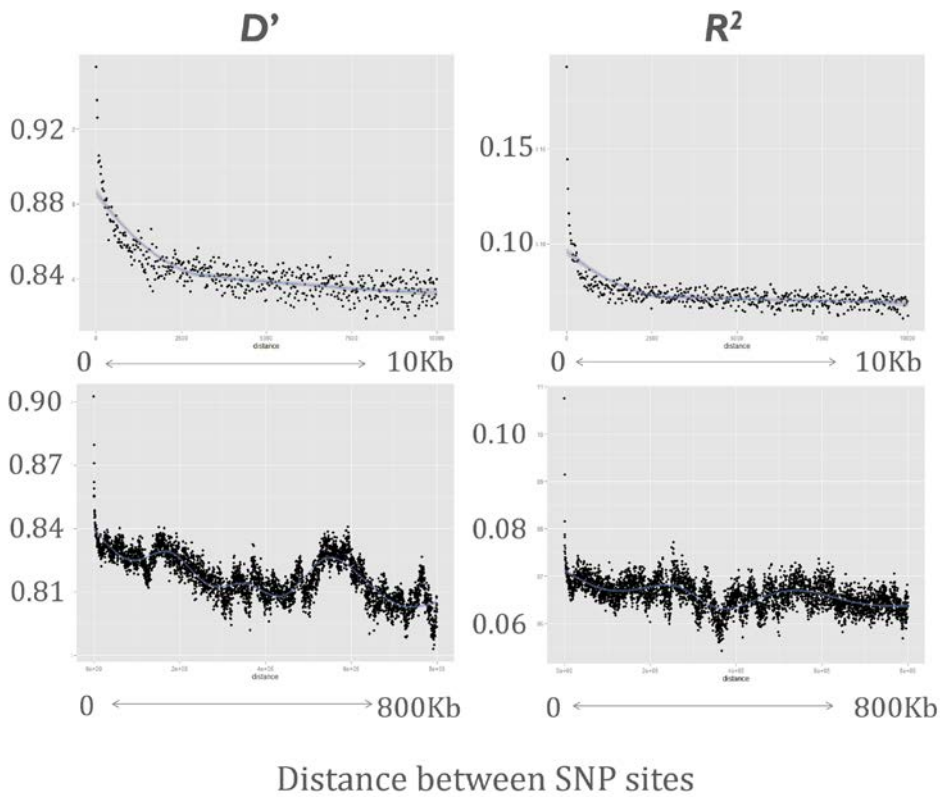

Figure 27. LD-decay within *E. coli* genomes at short and long physical distance.

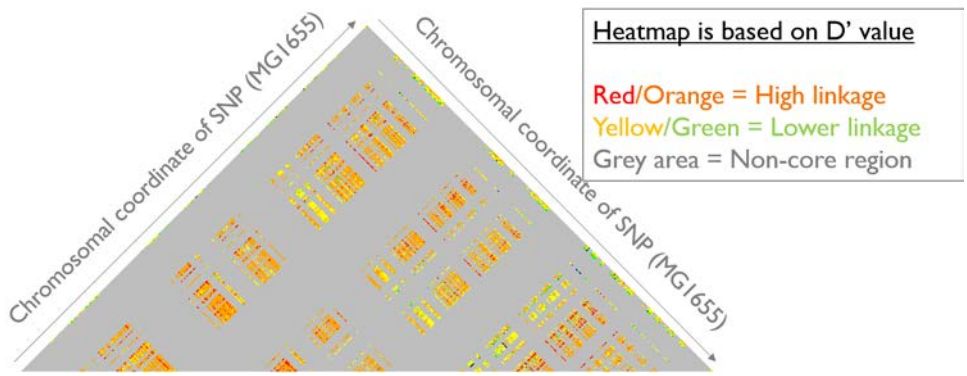

Figure 28. Chromosome-wide linkage heat-map based on *D'* values. High *D'* values between the SNP sites separated by long distances indicated the presence of long range linkage throughout the core-genome of *E. coli* (i.e. the clonal frame phenomenon).

2.3.9. Comparison of *E. coli* pan-genome properties and phylogenetic structure with those of other bacterial species

To provide broader perspectives to the characterization of *E. coli* pan-genome made in this chapter, pan-genomes of diverse other bacterial species were analyzed and compared. The basic question was to know if the observations made for *E. coli* pan-genome was universally applied to diverse bacterial species. Comparison between bacterial species that have different ecological niche was expected to offer some insights about the relationship between pan-genome property and ecological property of bacterial species. The list of species analyzed here includes commensal enteric bacteria, opportunistic pathogens, obligate pathogens, intracellular pathogens, and free-living environmental bacteria.

Gene frequency distribution obtained for the pan-genomes of the species analyzed were shown in **Fig. 29**. None of the examined species deviated significantly from the general rule of 'U-shaped' distribution. Another general feature was that the slope on the left side of distribution was less steep than the slope on the right side of the distribution. In other word, the core-genomes generally tended to have less "near-core" genes, while the singleton genes have more "near-singleton" genes. Evolutionary rates of pan-genome growth was analyzed for multiple species dataset. Two questions were under this analysis. First, can we generally obtain the linear relationship between pan-genome size and the phylogenetic diversity occupied by the strains? Second, which species exhibit steeper slopes, so that the species had higher rate of pan-genome expansion normalized by core-genome diversification? In the results, shown in **Fig. 30**, linear relationship was not always obtained in all

species. Still, 28 of the 39 species studied showed linear shapes. Among the 11 species that did not show linear relationship, 4 species (*H. pylori, C. jejuni, V. parahaemolyticus* and *B. mallei*) had almost linear correlation but their data points were slightly curved in a way that as x value became larger the slope decreased slightly. Such a pattern is likely to be an indicator of closed pan-genome. The other 7 species displayed multiple lines in the plot: *C. trachomatis, L. pneumophila, E. faecalis, E. aerogenes, M. abscessus, C. coli* and *P. acnes*. Negative correlation between the end-point total branch length of phylogenetic tree and the slope of the plot was observed in the comparative analysis of *E. coli* groups. In the multi-species comparison, such a trend was not obvious. The slopes of linear regressions shown in the **Fig. 30** were compared to illustrate relative rates of pan-genome expansion in these species (**Fig. 31**). *E. coli* was placed in intermediate rank. Extremely slow slopes were observed in the pan-genomes of *H. pylori, H. influenzae, C. trachomatis* and *C. jejuni*. On the other hand extremely rapid slopes were observed in the pan-genomes of *B. fragilis, Y. pestis* and *Y. pseudotuberculosis*. Based on the results provided in this section, *E. coli* pan-genome seemed to have average level of pan-genome growth rate among bacteria.

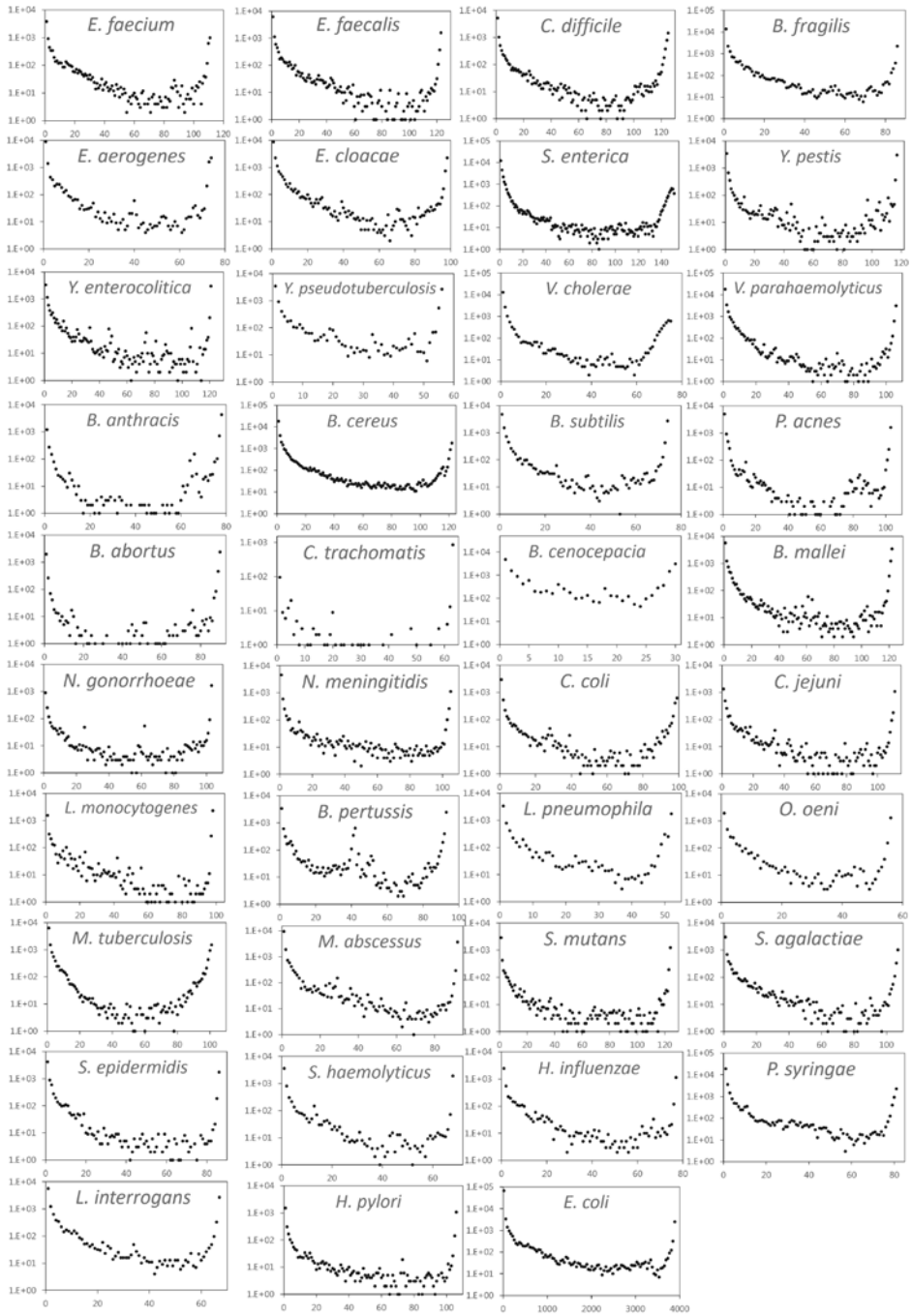

Figure 29. Gene frequency distribution of diverse bacterial species.

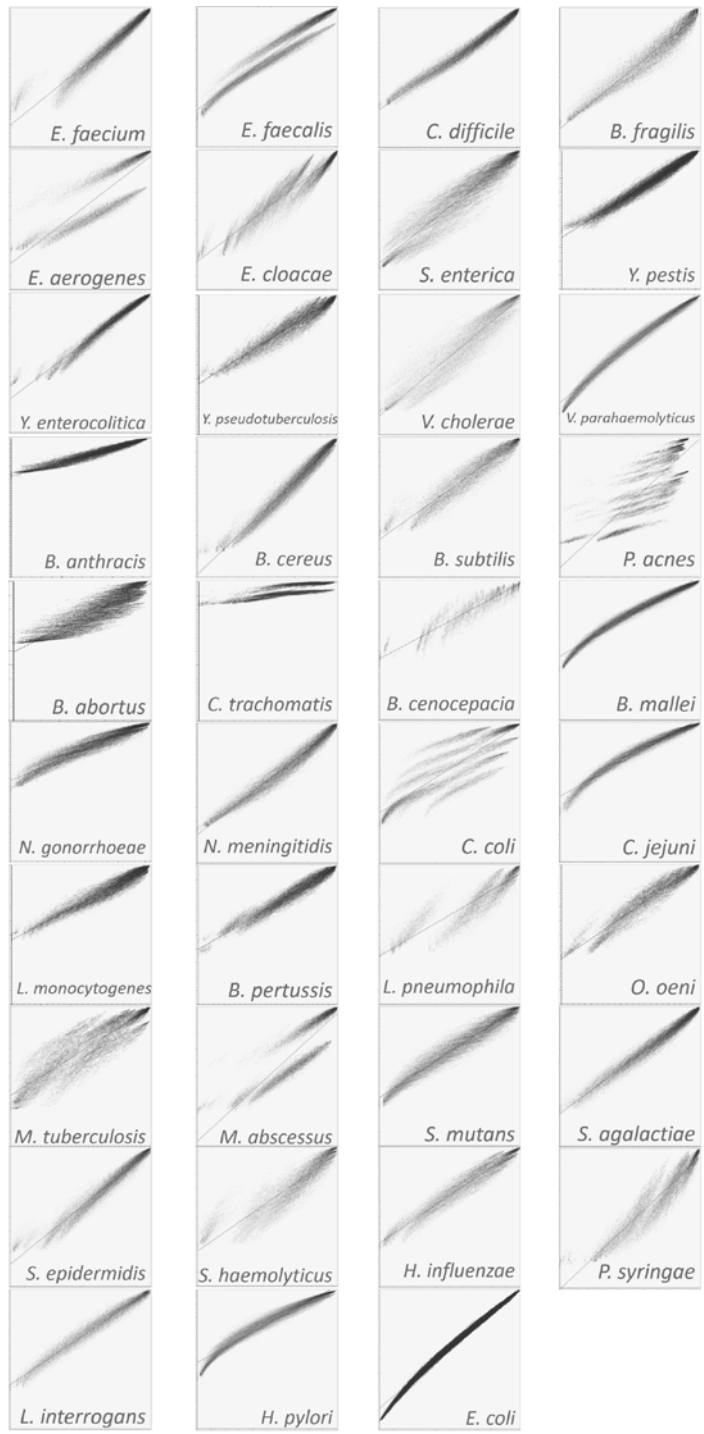

Figure 30. Pan-genome growth curves normalized by phylogenetic diversity of the strains, for 39 diverse bacterial species.

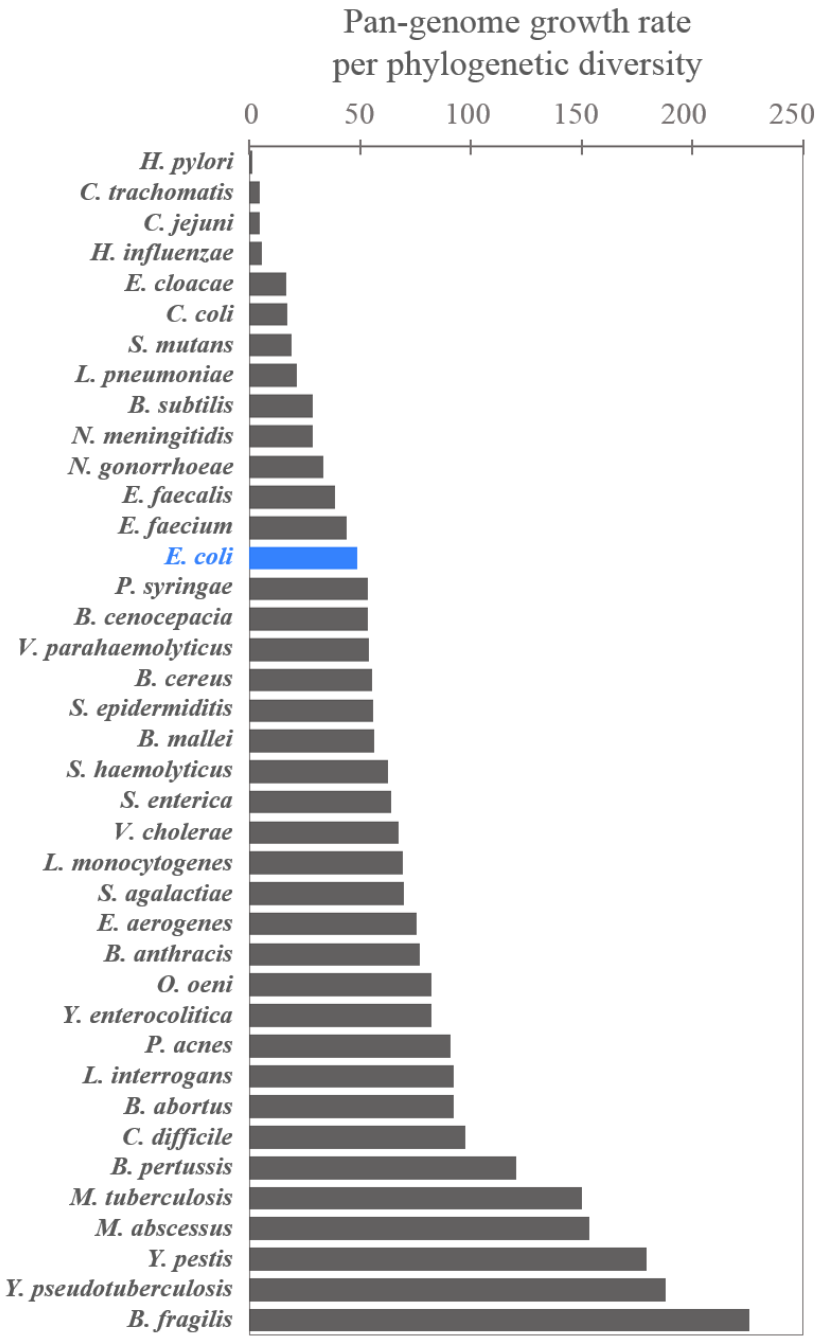

Figure 31. Relative ratio between pan-genome growth and phylogenetic diversity of diverse bacterial species. Species were ranked by the estimated ratio. *E. coli* was marked by blue color.

Core genome sequence diversity of *E. coli* was compared with that of diverse other bacterial species. Generally all species showed unimodal distribution of nucleotide diversity of core genes. The mode of the distribution seemed to be a good indicator of the genetic diversity within species. Among the 39 species analyzed, the degree of intra-specific genetic diversity varied greatly (**Fig. 32**). *B. abortus, B. anthracis* and *Y. pestis* displayed the lowest sequence diversity. Sequence diversity of *E. coli* core-genome was comparable with that of *V. cholerae, C. jejuni, S. enterica, Y. enterocolitica, E. aerogenes, V. parahaemolyticus* and *M. abscessus*. Extremely great sequence diversity was observed in the core-genomes of *H. pylori, H. influenzae, B. cereus, E. cloacae* and *P. syringae*. Based on the modal nucleotide diversity of core genes, the intra-species genetic diversity of these extremely diverse species were 3-5 times greater than that of *E. coli*.

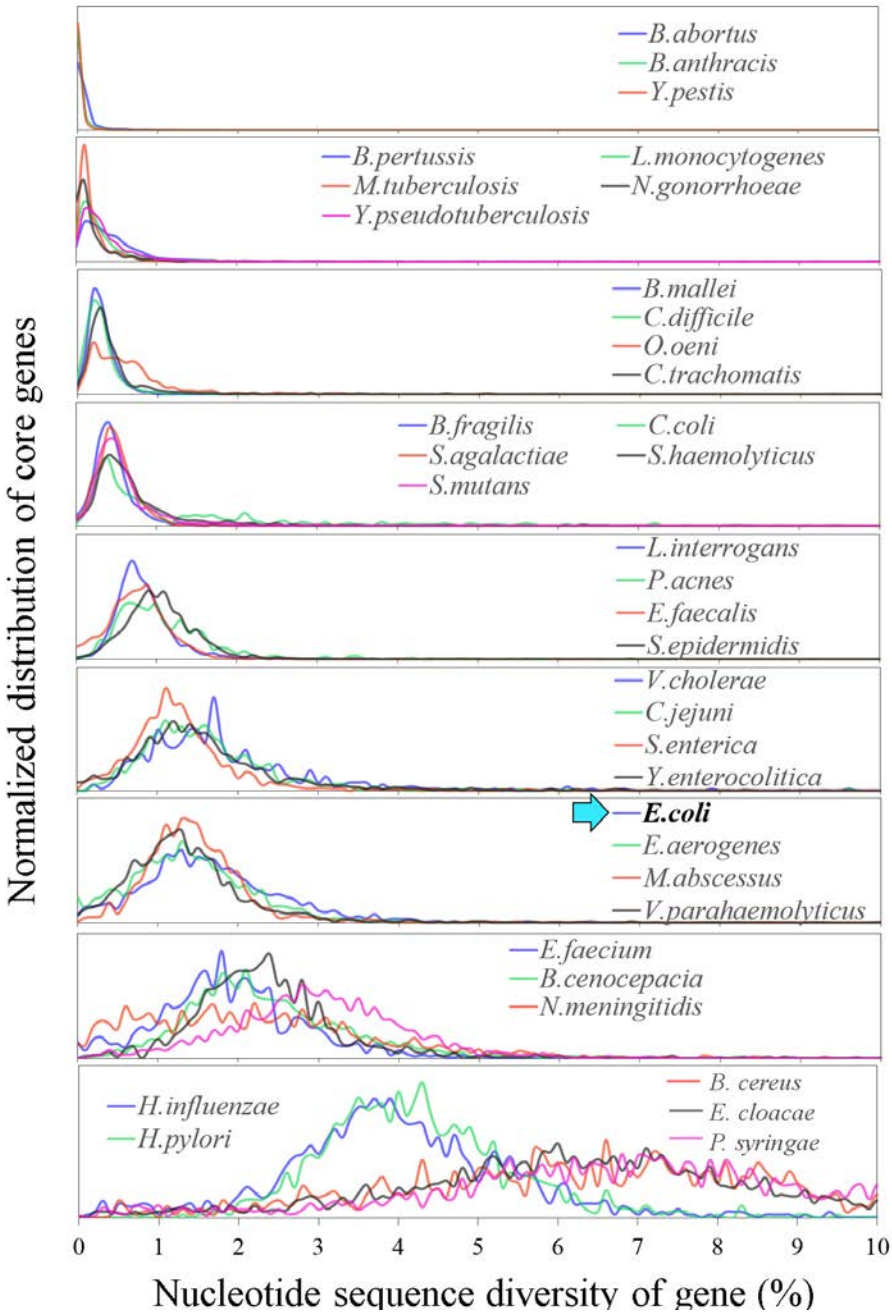

Figure 32. Nucleotide diversity of core genes in diverse bacterial species. The species were categorized by the modal nucleotide diversity of core-genome. Position of *E. coli* was marked by an arrow.

2.4. Discussion

Various aspects of the genomic diversity of *E. coli* were explored in this chapter. In the emerging picture of *E. coli* genomic diversity, the presence of phylogenetic structure within the species and the openness of their pan-genome were noteworthy. Phylogenetic structure of *E. coli* could be described by 7 major monophyletic clades of *E. coli* strains embedded by 4 monophyletic clades of *Shigella* strains and 6 miscellaneous clades of the strains that branched outside of the major clades. Clustering of the strains into phylogenetic groups were not in conflict with the clustering of strains by population structure inference based on SNP data. Presence of phylogenetic structure within a species is an interesting phenomenon. A strict concept of species, as a basic unit of evolution, would not allow presence of phylogenetic structure inside a species. Bacterial species concepts were proposed to be more flexible and ambiguous, to take into account the complexity of the roles played by mutation, recombination and natural selection as cohesive and divergent forces in bacterial populations (Fraser, et al. 2007; Achtman and Wagner 2008). The presence of *E. coli* phylogenetic groups has been interpreted as a sign of on-going speciation by some researchers based on lower inter-group recombination rates and higher intra-group recombination rates (Didelot, et al. 2012). In this study, genome-wide LD-decay was shown to be weaker in each phylogenetic group, than in the species *E. coli* as whole. That could be interpreted as elevated linkage within each phylogenetic group, and in turn, as shifting toward clonal genetics within each phylogenetic group. Branching patterns in the phylogeny also supported the speculation that genetics of *E. coli* has shifted from sexual to clonal genetics. Branching patterns of the phylogenetic trees of the strains of extremely sexual

bacteria were characterized by dominance of long branches that radiated near the root, while that of extremely clonal bacteria were characterized by concatenation of short branching events. Phylogeny of *E. coli* strains was composed of radiation of long branches in the basal area and the series of short branches near the tip area within each phylogenetic group. Evidences of shifting toward clonal evolution within each phylogenetic group will be discussed more thoroughly in the chapter 3, following the estimations of recombination rates.

Statistical analysis of the pan-genome growth curve based on exponential growth function indicated that the pan-genome growth will not stop as the number of genomes grow. In other word, open pan-genome was observed for *E. coli.* Openness of the pan-genome could mean that the pool of genes from which *E. coli* population has a chance of genomic influx is unmeasurably large. That pool of genes are possibly consisted by the entire gene diversity of bacteria living in the environments that are habitable to *E. coli.* One interesting feature in the pan-genome growth curve was that singleton genes which occurred only once (in a single strain) contributed to about 35% of the pan-genome size and that proportion was not decreasing significantly after the accumulation of 3,909 strains. Genes that occurred just once may not represent the functional components that have been actually integrated into the systems biology of *E. coli* cells. When the genes that occurred in 1% or more strains were taken into account, the number of such genes converged quickly and maintained around 15,400 genes as the number of strains increased. Therefore, the growth of pan-genome size were driven by the genes that occurred in less than 1% of the strains.

In the pan-genome growth curve analysis, the size of pan-genome was modeled as a variable that depends on the number of strains. A limitation of that model was that the extent of diversity within the sampled strains was not considered as a factor. In this study, the number of strains was replaced by the phylogenetic diversity of strains calculated as the sum of branch length occupied by the strains in the core-genome phylogeny. While the pan-genome size was dependent on the number of strains by exponential function, the phylogenetic diversity of strains and the pan-genome size showed a linear correlation. In the linear function, the slope of the function could be interpreted as the rate of the gaining of pan-genome size per unit increase in the phylogenetic diversity of the sampled strains. This rate, the relative rate of gene repertoire expansion over sequence diversification, could be used as a value characterizing genomic evolutionary property of a bacterial species. According to this measure each phylogenetic group within *E. coli* showed higher relative rate of gene repertoire expansion over sequence diversification. Inflation of the rate could be due to increased rates of gene gain, or, decreased rate of sequence diversification. With several reasons the latter explanation, the decreased rate of sequence diversification, seemed to be plausible. First, gene frequency distributions of the pan-genome of phylogenetic groups were generally overlapping. Second, from direct comparison between the pairwise divergence of gene contents and the pairwise divergence of core-genome sequences it seemed that while gene contents can be differ significantly without sequence divergence, sequence divergence does not occur without gene content divergence. Third, the relative rate of pan-genome growth over phylogenetic diversity was negatively correlated with the total phylogenetic diversity of the group. The negative correlation trend was much weaker when randomly sampled subsets of strains were analyzed instead of phylogenetic

groups. The two groups that showed the highest prevalence of intra-group recombination of core genes deviated from that negatively correlated trend toward the values obtained from random subsets. Slower sequence diversification in more clonal populations can be expected by increased tendency of genome-wide selective sweep (which purges the diversity) because of increased genome-wide linkage. To argue that the selective sweep played its role to slow down the core-genome sequence diversification within the phylogenetic groups due to their clonality, evidences should be added from analysis of natural selection and recombination rates. In the chapter 3, such evidences will be discussed further.

Sequence diversity (polymorphism and divergence) within orthologous genes comprise one aspect of genomic diversity within the species. Genes of high intra-population gene frequency in *E. coli* showed a unimodal distribution of nucleotide diversity. Low frequency genes (rare genes) showed more dispersed distribution of sequence diversity. The mode of sequence diversity of core genes seem to be a good indicator of genetic diversity/polymorphism within the species because of their unimodal distribution. The modal sequence diversity of *E. coli* core genes was 1.3%, which was comparable with that of *C. jejuni, E. aerogenes, M. abscessus, S. enterica, V. cholera, V. parahaemolyticus* and *Y. parahaemolyticus*. Gene content variation comprises another aspect of genomic diversity within the species. When Bray-Curtis dissimilarity statistic was calculated for all pairs of strains, the dissimilarity values were distributed in a largely unimodal shape, with the peak at 0.23. Based on this value, if you peak any two *E. coli* genomes and count the genes that are shared and not shared between the two strains, most frequently you will observe that 23% of the genes in the strain are not shared between the two

strains. Variations of gene order, or synteny, can be also considered as an aspect of genomic diversity. Analysis of synteny within the core-genome of *E. coli* revealed that the synteny was remarkably well-conserved along the core-genome even in the presence of severe gene content divergence. It is not clear at this point, if the conservation of gene order resulted from selective constraint that acted upon the variation of gene order. Along with the observation of gene content divergence that happened independent of core-genome sequence divergence, the conservation of stable core gene order implies that the evolution of core-genome sequences is uncoupled from the evolution of gene repertoire. Patterns of gene presence/absence in *E. coli* strains were analyzed in the frame of phylogeny, and the genes that are specific to and conserved in each phylogenetic group could have been identified. Genes specific and conserved to a species or a group could be regarded as diagnostic genes. Group B1 and A, which are composed of strains that show high ecological diversity had little diagnostic genes. Group SII, SIII and E had a large number of diagnostic genes (about 150 genes). The latter groups are exclusively composed of pathogenic strains. Interestingly, the number of genes that are diagnostic to *E. coli* species (specific core genes of *E. coli*) was 40 and it was smaller than that of the pathogenic groups SII, SIII, SIV, E and larger than that of other groups composed of ecologically diverse strains. Such observation implied that the evolution of pathogenicity in *E. coli* was at least partially driven by acquisition of genes.

Study of intra-specific genomic diversity has an explorative nature like other area of biodiversity studies. Despite of exceptionally rich genome data available for *E. coli*, a comprehensive and systematic study of ecologically adaptive evolution of the genomes of this bacterial species was not feasible. First, the natural habitats of

E. coli are much more diverse than what is represented in the current genome dataset. To investigate adaptive genome evolution related to host range we need more genome sequences from the strains isolated from diverse habitats. Pathogenicity evolution also requires more strains that are designated to be “commensal” strains. The number of strains that have been declared to be “commensal” was found to be surprisingly small. Actually, the strains newly sequenced in this study that were isolated from river water (thus, environmental strain) were not able to be designated as “commensal” because the infection phenotypes of these strains were not tested. A caution had to be taken because some of these environmental isolates shared the highest genomic similarity with known pathogenic strains. Therefore, from ecological and evolutionary genomics perspective, the richness of genome data of *E. coli* should be followed by strain-level characterization of phenotypic properties and an effort should be made to increase the genome sequences of non-human isolates.

CHAPTER 3

Characterization of microevolutionary processes that mediated genomic diversification of *E. coli*

3.1. Introduction

Microevolution refers to the evolution that take place inside the species. The process that have shaped the standing genomic diversity of a species and the processes that govern the genetic diversity of a species all belong to the area of microevolution. In the previous chapter the diversity represented in *E. coli* genomes were explored. In the chapter 3 the impacts and roles of various types of microevolutionary mechanisms in the shaping of current genomic diversity *of E. coli* were addressed.

Clonality vs. sexuality of bacterial species has been a long-standing and exciting theoretical problem in bacterial genetics. As bacteria are asexually reproducing organisms their microevolutionary process is clonal by default. Sexual process of bacterial microevolution, accompanying exchange of alleles between individuals in the species, is realized through homologous recombination of DNA segments. Historically, before the use of molecular genotyping techniques researchers guessed that *E. coli* population had a clonal nature. At that time one of the strongest evidence for clonality was the patterns observed in serotypes. Combining the number of types observed for each of O-, H- and K-antigens there were 900,000 possible combinations of O:H:K serotypes, but actually detected number of serotypes were not that many. In 1970s, some studies based on multilocus enzyme electrophoresis technique first recognized the importance of recombination in *E. coli* population. In 1990s sequencing-based analyses were carried out and the researchers began to suggest that the impact of recombination was more important than mutation in *E. coli* (Dykhuizen and Green 1991; Guttman and Dykhuizen 1994;

Smith 1999). In a paper that became a classic reference (Wirth, et al. 2006) MLST data was used to analyze the recombination, mutation and population structure of *E. coli*. In the article, the conclusion was a hypothesis that commensal strains maintain low recombination rate while virulent strains experience higher recombination rates. In more recent studies (Leopold, et al. 2011) a hypothesis was proposed that *E. coli* were freely recombining in the past but currently they have entered a recombinational dormancy, based on the analysis of MLST data. In 2012, whole genome analysis using 27 *E. coli* strains concluded that group A+B1, B2 and E are undergoing sexual isolation (Didelot, et al. 2012). The most recent study using whole genome data concluded almost equal impact (r/m=0.92) of recombination and mutation in *E. coli* genome evolution (Bobay, et al. 2015). Still there has been no large scale analysis of whole genome data that addressed the impact of recombination.

Horizontal gene transfer is a microevolutionary process that is uniquely important in the populations of bacterial species. Rate of horizontal gene transfer in bacteria was reported to be so fast that several genes are lost and gained while a single nonsynonymous substitution is fixed in a conserved gene (Puigbò, et al. 2014). Our ability to detect and source-track the horizontally transferred genes in the genome depend on the size and quality genome sequence database that covers entire taxonomic diversity of microorganisms. In chapter 2 we have observed that the significant proportion of the pan-genomes of bacterial species is composed of phyletically rare genes. Even considering the possibility of birth of new genes in the species and the possibility of misprediction of protein-coding sequences, horizontal acquisition of xenologous genes seemed to have a pervasive impact on genomic diversification within bacterial species.

A final important aspect of microevolutionary process is the action of natural selection. Natural selection could act in three direction. It can suppress the diversification of sequences, by what is called a negative (or purifying) selection. It can promote the diversification of the sequences by giving accelerated fixation rate to the newly arisen mutation, when it happens in the form of positive (or diversifying) selection. Lastly, by not having an impact, the sequences are set free to diversify neutrally under genetic drift. Relative contribution of natural selection in the shaping of sequence diversity have been at the center of long-standing debate between neutralists and selectionists. Unfortunately, action of natural selection in natural populations of bacteria is considered to be difficult to study because the genomic loci are clonally linked in bacteria (Corbett-Detig, et al. 2015).

In the current chapter the impact of homologous recombination, the impact of gene gain, loss and transfer, and the impact of natural selection in the diversification of *E. coli* genomes were evaluated. Specifically, the following questions oriented the analyses:

i. How pervasive was the impact of homologous recombination during *E. coli* genome evolution?

ii. Did recombination happen evenly across the core-genome and along the history?

iii. Was recombination biased by phylogenetic structure within the *E. coli*?

iv. How many gene gain and loss happened during *E. coli* genome evolution?

v. What is the origin of singleton genes in the *E. coli* pan-genome?

vi. How does impact of natural selection differ in the core genes and dispensable genes, and what is the direction of natural selection in *E. coli* genomes?

3.2. Materials and methods

3.2.1. Genome dataset

Available statistical tools for genome-wide inference of recombination history were not appropriate to be applied to the entire genome dataset used in the chapter 2, because computational load for BratNextGen and ClonalFrameML for more than 1,000 genomes was unaffordable. As a result, the genome dataset was reduced by selecting the representative strains. Based on the phylogenetic tree, 325 strains were manually selected to evenly cover the clades within *E. coli*. Dataset used for gene-by-gene determination of the presence of recombination was the same dataset that was used in the pan-genome analysis described in the chapter 2. Gene gain and loss analysis was performed with 1,897 *E. coli* genomes that were available at the time of analysis (March 2015). The most widely used tool for calculation of dN and dS from the input sequence alignment and phylogenetic tree, the codeml program in the PAML package, was somehow not able to process large number of sequences. By that reason, 155 representative *E. coli* strains were again selected based on the manual inspection of phylogenetic tree and used for calculation of intra-specific dN and dS. For gene-by-gene detection of recombination in the sequence alignment, all 3,909 *E. coli*/*Shigella* genomes that were described in the chapter 2 were used because PhiPack test of recombination performed at reasonable speed even for large number of sequences.

3.2.2. Analysis of homologous recombination events

Genome-wide inference of recombination history was performed by two popular statistical methods, one provided by BratNextGen software (Marttinen, et al. 2012) and the other provided by ClonalFrameML software (Didelot and Wilson 2015). BratNextGen takes the whole genome alignment as an input. Then it divides the genome alignment into 5 kb blocks, merges the neighboring blocks if one of the block does not contain enough SNP sites (at cutoff 20), and performs clustering analysis for each block. Each resulting block is then subjected to clustering based on the SNP matrix. Based on the frequency that each pair of strains were found in the same cluster, a PSA (proportions of shared ancestry) tree is estimated and used to create the initial clustering of strains. Based on the initial clustering, MCMC-like iteration is then run to learn the parameters such as mutation rate and recombination rate. Then probability of recombination is inferred by the Bayesian change-point clustering model that allow for different clustering in different genomic regions. The origin of recombination is modeled to be variable between genomic regions. In our analysis the initial clustering to begin the analysis was set to be 45 clusters, so that each phylogenetic group contained at least 2 clusters. As an exception, all group E strains were so closely related in the PSA tree that it was unable to set more than 1 groups within the group E. The learning step was repeated for 20 iterations. Recombination inference was run in 20 replications and the significance limit of 0.05 was used. The output table contained the coordinates of the start-end points of the recombined segments in each input genome and ancestral genome. The inferred origin of the segment was also written in the output. This output table was parsed to calculate the basic statistics of recombination and the amount of gene flow between the phylogenetic groups.

ClonalFrameML basically takes a whole genome alignment and a phylogenetic tree and infer the recombination history based on the assumption that the given phylogeny is the clonal genealogy of the input genomes. To use as a phylogenetic tree input for ClonalFrameML, a maximum-likelihood phylogenetic tree of the 325 strains in the dataset was reconstructed by PhyML (Guindon, et al. 2010) using the concatenated core-genome alignment under the general time-reversible model. The parameters such as median recombination rate, median recombination tract length, and the branch lengths in the input clonal genealogy were set to be estimated by the Baum-Welch Expectation Maximization algorithm for each branch in the genealogy. The resulting output contained the begin-end coordinates of the imported segments in each strain and the ancestral genomes, as well as the branch-specific Expectation-Maximization values of R/theta (relative ratio of population-level recombination rate and mutation rate) and mean length of imported DNA segments. The output was parsed to obtain the relationship between the node-height and R/theta, and the distribution of R/theta in the extant branches.

Gene-by-gene detection of recombination was performed by the PHI test of PhiPack package, as described in the previous chapter. In the chapter 2, the result of PHI test was used to filter out the recombinant genes from phylogenetic analysis. On the other hand, in this chapter, the result of PHI test was used to calculate the prevalence (%) of genes that showed a significant sign of recombination.

3.2.3. Analysis of gene gain and loss history and tracking the origins of the singleton genes in *E. coli* pan-genome

Gene presence/absence matrix for 1,897 *E. coli* strains and the maximum-likelihood phylogeny of core-SNPs of the strains were used as input to run the GLOOME version VR01 (Cohen, et al. 2010). GLOOME is a model-based approach and that search for the optimal scenario of gain and loss using stochastic mapping approach. The model provided by GLOOME had flexibility to allow gain and loss rates and the gain/loss ratio to be variable among the genes. However, in order to process the large dataset used in this analysis, the model parameters were set to have low flexibility. Distribution of the rates among the sites was selected to have gamma distribution, and the number of gain rate categories was set as 3, the number of loss rate categories as 3, and the number of gain/loss ratio categories as 3. The optimization level was set to be “low”.

Protein sequence homologs of the singleton genes in the pan-genome of *E. coli* were searched in the genome database that contained 44,140 prokaryotic genomes. A fasta file containing the representative amino acid sequences of the singleton genes was queried against the proteome of each genome in the database. Ublast tool implemented in the USEARCH 8 (Edgar 2010) was used for alignment searching. In this analysis, because the purpose was to find the recent gene transfer or recent gene duplication, there was no need to find the distant homologs that do not share high similarity and high length coverage with the query gene. Therefore the identify cut-off was set to be 70% and the query coverage cut-off was set to be 90% in the ublast search. Based on the presence/absence of non-self-hits in the *E.*

coli genomes and the presence/absence of hits in the genomes of other prokaryotes, the singleton genes were classified into 4 categories. The genes that had no hit in any prokaryotic genome were classified as true ORFans. The genes that only had hits in *E. coli* genomes were classified as putatively recently duplicated genes. The genes that only had hits in non-*E. coli* genomes were classified as putatively xenologous genes, which means that the gene were transferred from other species. Because the genes were found in only one strain of *E. coli*, putatively xenologous genes are not likely transferred from the *E. coli* strain to the matched strain. Rather, the presence of close homologs between 1 *E. coli* strain (out of 3,909) and the strains of distant species is likely due to (i) transfer of the gene from that species, or (ii) sharing of same environmental gene pool. The genes that had hits in both *E. coli* and non-*E. coli* genomes required phylogenetic analysis to determine if the gene was xenologous or not.

3.2.4. Analysis of dN/dS ratio

DNA sequences of each orthologous gene cluster in the pan-genome of *E. coli* was aligned codon-by-codon. Briefly the DNA sequences were translated to amino acid sequences and amino acid sequences were aligned by MAFFT. The aligned amino acid sequences were used as templates to generate DNA alignment. The alignment files were converted to nexus files for coldeml analysis. The core-genome phylogeny was also prepared for codeml input. For phylogenetic reconstruction of 155 genomes by ML method, PhyML was run with HKY model of substitution and Gamma distributed rates. For each alignment the codeml program of PAML4 (Yang 2007) was run with –runmode option 0, to estimate the rates of synonymous and nonsynonymous substitutions in the codons. Under runmode 0, the codeml program used the input phylogenetic tree as a part of evolutionary model. Model parameter was set to 2, so that branches could have variations of rates. The output file contained a matrix of pairwise dN, dS, and dN/dS estimated by Nei & Gojobori 1986 method (Nei and Gojobori 1986). For each gene, pairwise estimates were averaged to make a dataset-average value of dN, dS and dN/dS. The same analyses was repeated using the input sequence alignments that contained only the members of one specific phylogenetic group, to compare the level of selection within the species and within the phylogenetic groups. The resulting dN, dS and dN/dS values were also summarized separately for gene frequency categories. Gene frequency categories were defined to represent phyletically rare genes (2.5-9% of strains), outer “shell” genes (9-75%), inner “shell” genes (75-97%), near-core genes (97-99%), and the core genes (99-100%).

3.3. Results

3.3.1. Impact of homologous recombination in genomic evolution of *E. coli*

Concatenated alignment of the core-genome of 325 strains selected for homologous recombination analyses were used to run BRATNextGen and ClonalFrameML. The results of BRATNextGen analysis showed that the total number and mean length of the segments imported by recombination differed significantly between phylogenetic groups (**Fig. 33**). To quantify the collective impact of recombination to the genome sequences, the total length of imported sequences were analyzed. The result shown in the **Fig. 34** suggested that group B2, D and F have experienced larger impact of recombination than the other groups. Mean length of segments imported by recombination was 910 bp and each *E. coli* genome contains 130 imported segments on average. Collectively, 118 Kb of foreign segments were imported to each strain, on average. Since the median genome size of *E. coli* was calculated to be 5.12 Mb, the estimated 118 Kb corresponds to 2.3% of the genome. Distribution of recombined segments along the core-genome was inspected to determine if recombination events were spatially clustered along the core-genome. The result shown in **Fig. 35** indicated that recombination happened all over the core-genome. A closer look at the frequency of recombination events that affected the genes revealed some genes that experienced relatively high impacts of recombination. The chromosomal order of core genes in MG1655 was compared

with the number of recombination events that affected the gene's evolution per strain (**Fig. 36**). The recombination-hot genes were marked by red circles in the **Fig. 36**. The functions of the products of the detected recombination-hot genes were summarized in **Table 6**. To see if phylogenetically close strains tend to share more common segments imported by recombination, the phylogenetic tree of 325 strains analyzed was aligned to the heat-map of per-gene recombination event counts (**Fig. 37**). Vertical lines were frequently found in the heat map, meaning that closely related strains have frequently received the same segments. It should be noted that the dots, strain-specific recombination events were also abundant in the heat-map. Presence of many dots implies that even between closely related strains recombined segments were not always shared. To determine if gene flow by recombination was biased toward specific pairs of the phylogenetic groups, total length of gene flow from one group to another group was summarized in a matrix with color gradient (**Fig. 38**). Gene flow directions were not symmetric as exemplified by large gene flow from B1 to A and restricted gene flow from A to B1, and were not even for inter-group directions. Group B2, D and F have received a particularly great amount of DNA segments from the group B2. Group of B1 and C strains received the least gene flow from the other groups. In accordance with previous studies (Leopold, et al. 2011; Didelot, et al. 2012), at least it was able to confirm the presence of specificity in the inter-group gene flow activity.

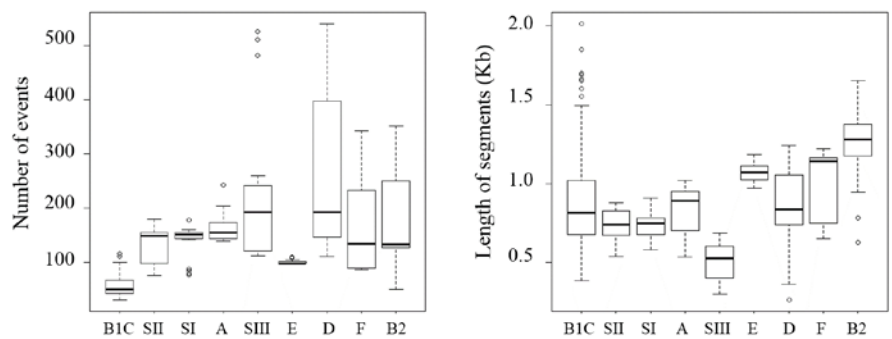

Figure 33. Properties of recombination events detected in the core-genome of *E. coli.* Left box plot is the group-by-group summary statistic of the number of recombination events that affected the genome of a strain. Right box plot is the summary statistic of mean length of recombination tracts detected.

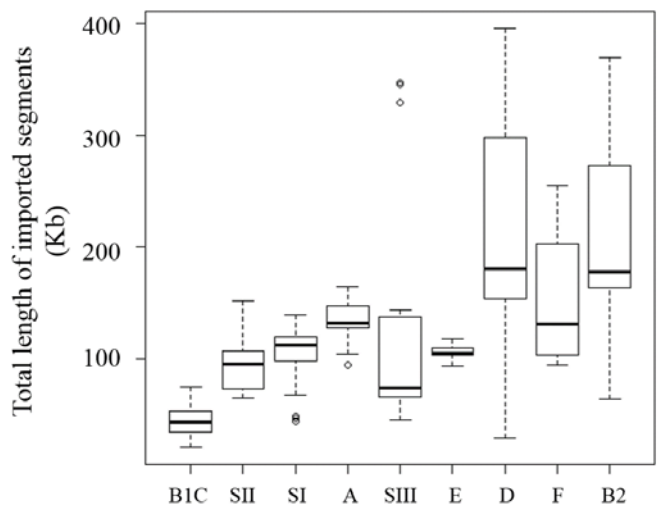

Figure 34. Total impact of recombination on the genomes of each phylogenetic group. The impact of recombination was quantified in terms of the total length of DNA segments imported to the genome by recombination.

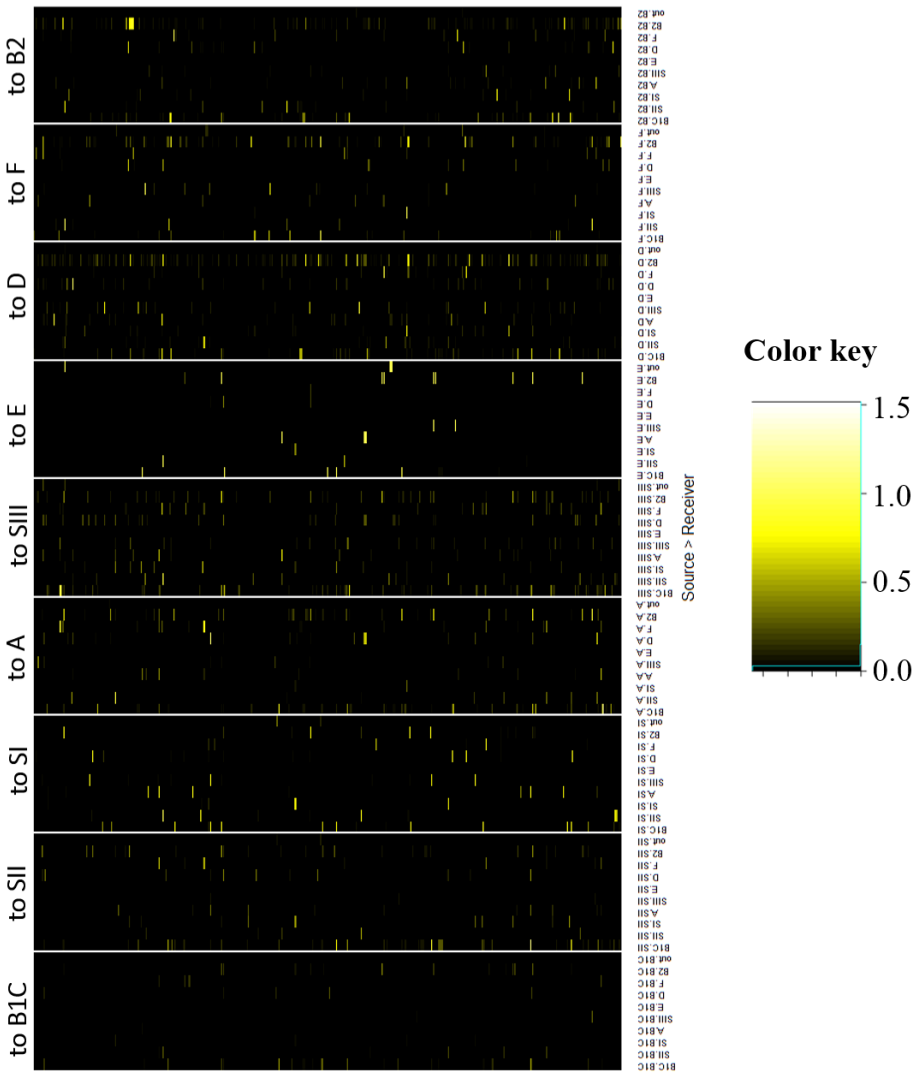

Figure 35. Heat-map of the number of recombination events per gene, per donor-recipient group combination. Core genes were ordered according to the chromosomal order. Bright cells represent the genes with high recombination rates.

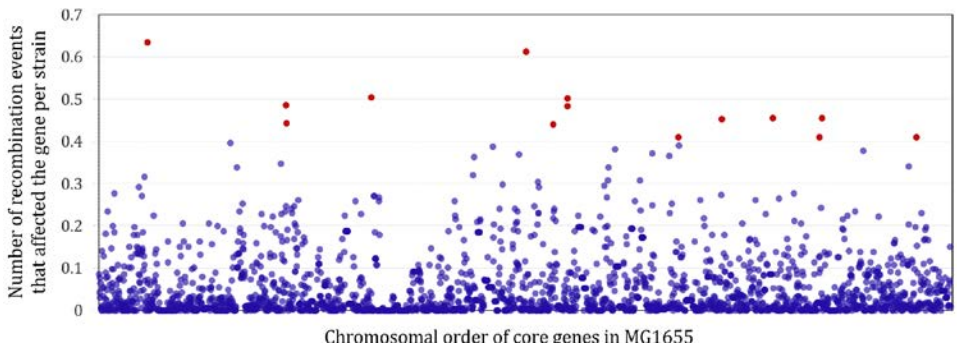

Figure 36. Distribution of recombination frequency per gene per strain. Each data point (circle) corresponds to a core gene. The core-genes were sorted on the x-axis according to their chromosomal order in MG1655. Recombination-hot genes were marked by red circles.

Table 6. Recombination-hot genes of *E. coli*.

Gene	Gene product	Recombination per strain
dnaE	DNA polymerase III alpha subunit	0.63
mnmC	Fused 5-methylaminomethyl-2-thiouridine-forming methyltransferase	0.61
hrpA	Putative ATP-dependent helicase	0.50
yfhL	Putative 4Fe-4S cluster-containing protein	0.50
ldtD	Murein *L,D*-transpeptidase	0.49
yfhH	Putative DNA-binding transcriptional regulator	0.48
bisC	Biotin sulfoxide reductase	0.46
wecC	UDP-*N*-acetyl-*D*-mannosaminuronic acid dehydrogenase	0.46
fusA	Protein chain elongation factor EF-G	0.45
ycbK	M15A protease-related periplasmic protein	0.44
ppx	Exopolyphosphatase	0.44
yhbU	U32 peptidase family protein	0.41
yjfP	Acyl-CoA esterase	0.41
rep	DNA helicase	0.41

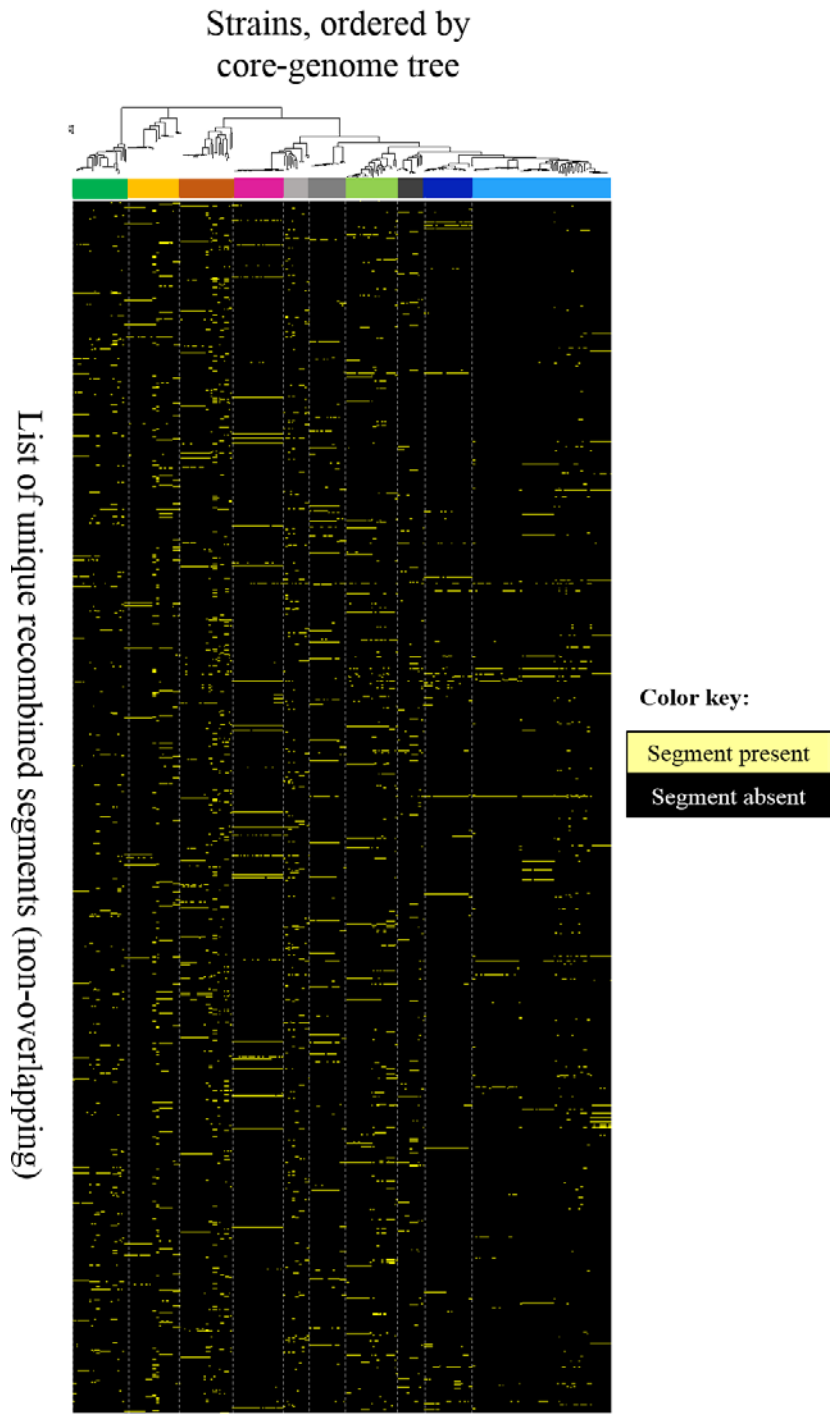

Figure 37. Co-occurrence of recombination in the strains phylogenetically closely related to each other.

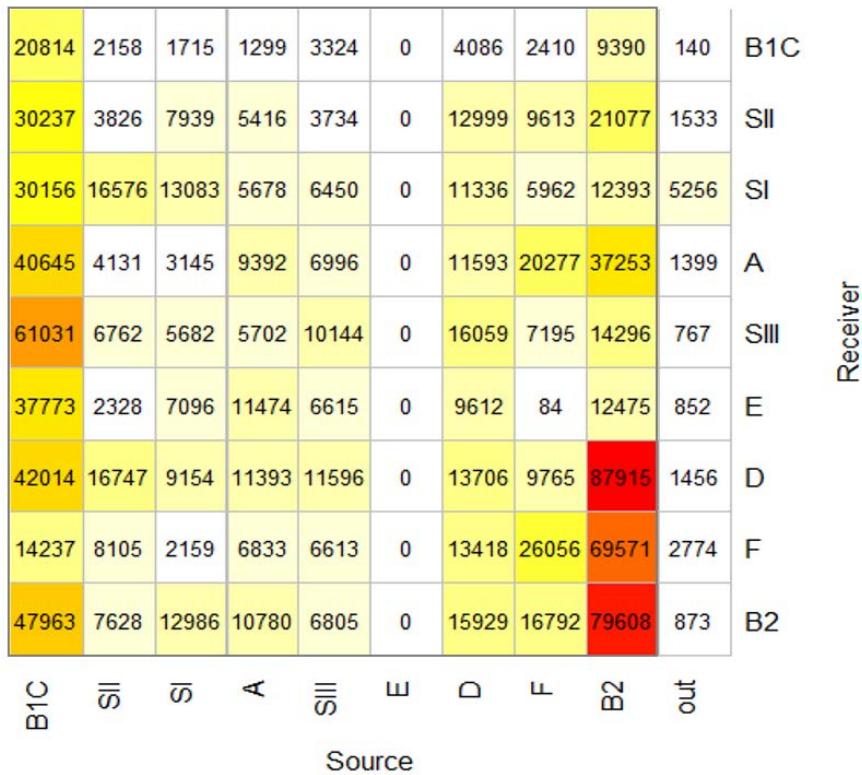

Figure 38. Amount of gene flow per source-receiver phylogenetic groups. Matrix values correspond to the average length (bp) of DNA segment flowed from the source group to the recipient genome per strain. Red color correspond to more active gene flow.

The results of ClonalFrameML analysis included the estimation of R/theta (recombination rate / mutation rate) for each internal and terminal branch of the core-genome phylogenetic tree. The estimated R/theta per each branch was compared with the height of the branch. Height of the branch was defined as the branch length needed to go from the closest terminal leaf to the mid-point of the branch. Branches with larger heights represented more ancestral lineages. **Fig. 39** shows the result of comparison between R/theta and height of the branches. In the ancestral branches estimated R/theta values were higher than 1, but in the recent branches the R/theta distributed largely below 1. Based on the result the relative impact of recombination have decreased temporally through the evolutionary history of *E. coli*. The recombination/mutation rate of extant *E. coli* population might be best described by the distribution of R/theta estimated in the terminal branches. In the terminal branches, R/theta values were distributed between the minimum value 0.11 and the maximum value 2.91. The median value was 0.73 and as shown in **Fig. 40** the R/theta values in the terminal branches were smaller than 1 for most majority of the branches. Interestingly the terminal branches with the smallest R/theta values were preferentially consisted of the branches leading to *Shigella* strains (from rank 1 to 8). The terminal branches with the largest R/theta values were preferentially consisted of the branches leading to group D strains (from rank 1 to 6).

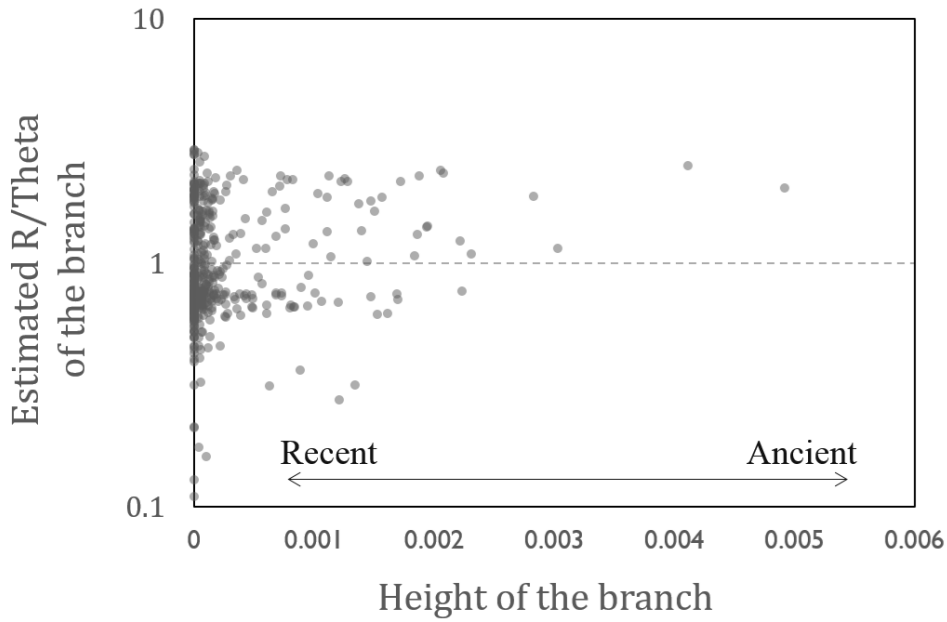

Figure 39. Decline of R/theta throughout the evolutionary history of *E. coli.* The horizontal dashed line represents the unity of recombination rate and mutation rate. Each data point (circle) corresponds to a single branch in the phylogenetic tree of the strains. Height of the branch is the distance from terminal node to the midpoint of the branch.

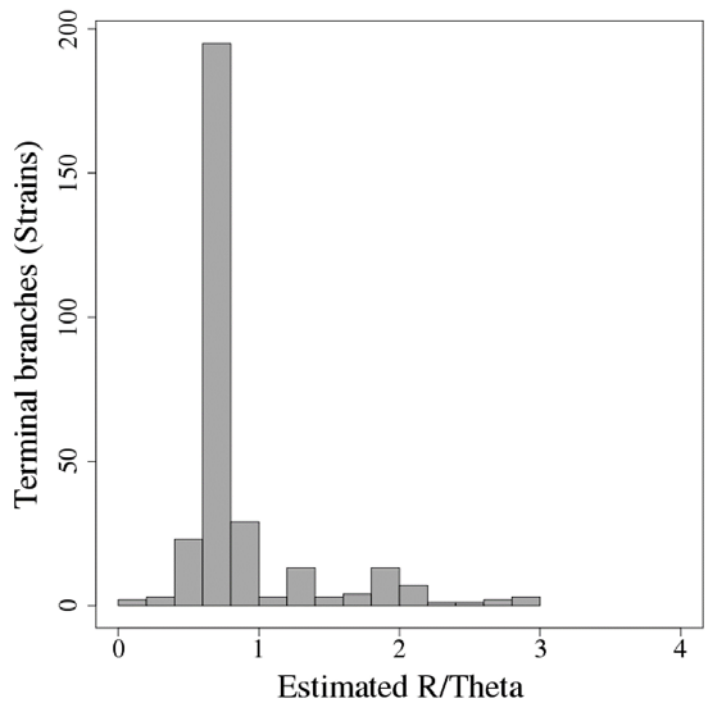

Figure 40. Values of R/theta estimated in the terminal branches.

Presence of recombination was tested for each gene sequence alignment for all genes in the pan-genome. Pan-genome was divided into five gene frequency categories. For each gene frequency category the proportions of recombination-detected genes were calculated and compared, to test whether the impact of recombination was more pervasive in the core genes than in the non-core genes. Five out of six species shared the same pattern. Proportion of recombined genes was highest in non-core genes (**Fig. 41**). Focusing on the core genes, proportions of recombined and non-recombined genes in the core-genome were estimated for 39 bacterial species (**Fig. 42)**. Species like *H. influenzae, H. pylori* and *N. meningitides* were extremely sexual, having recombination history in around 80% of their core genes. Extremely clonal bacterial species were *B. anthracis, Y. pestis, B. abortus* and *C. trachomatis* and less than 5% of the core genes of these species were detected as recombinants. In *E. coli* core-genome, the proportion of recombined genes was 54.8%. According to this result, *E. coli* not biased toward either ends of extremely clonal side or extremely sexual side. The correlation between the prevalence of recombination in the core genes of each species and the overall genetic diversity of species at sequence-level was tested. It seemed that bacterial species generally have larger genetic diversity when the recombination was more prevalent (**Fig. 43**). Exceptions from that general trend were *P. syringae, B. cereus* and *E. cloacae,* all of which had relatively low prevalence of recombined genes but exhibited very large nucleotide diversity. All of these 3 species were composed of two or more subspecies, and may had to be considered as the combinations of multiple species. Therefore the general trend was still valid, and none of highly clonal bacterial species has not been found to achieve significant genetic diversity.

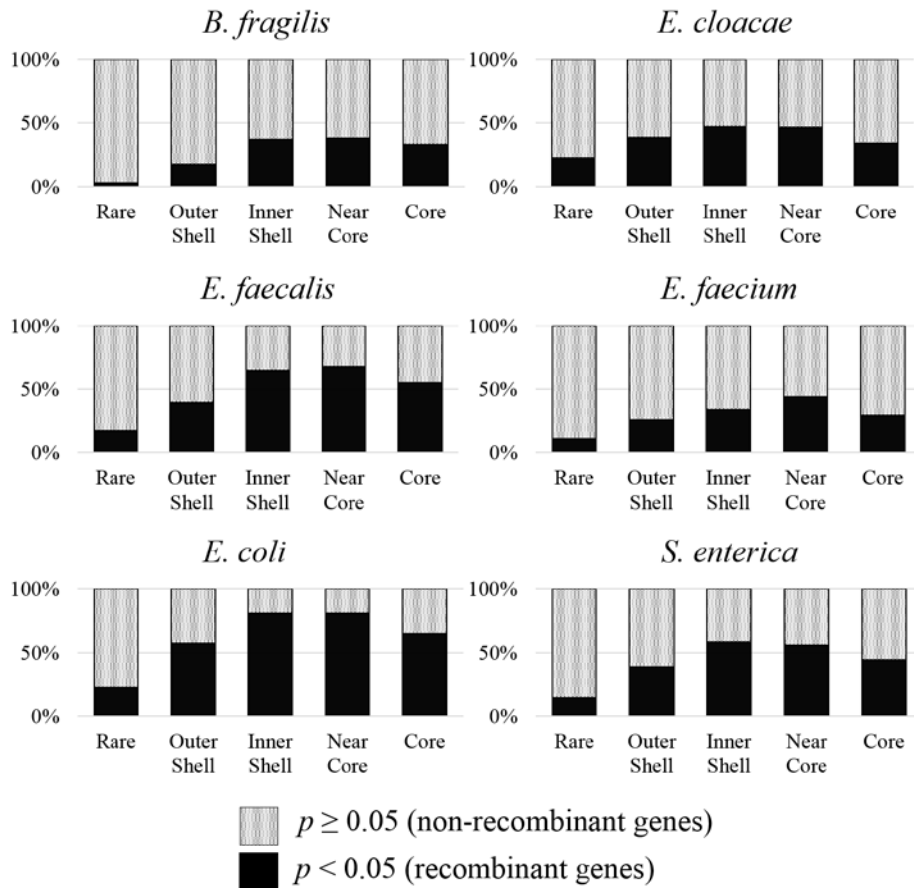

Figure 41. Prevalence of recombination in the genes compared by the gene frequency categories.

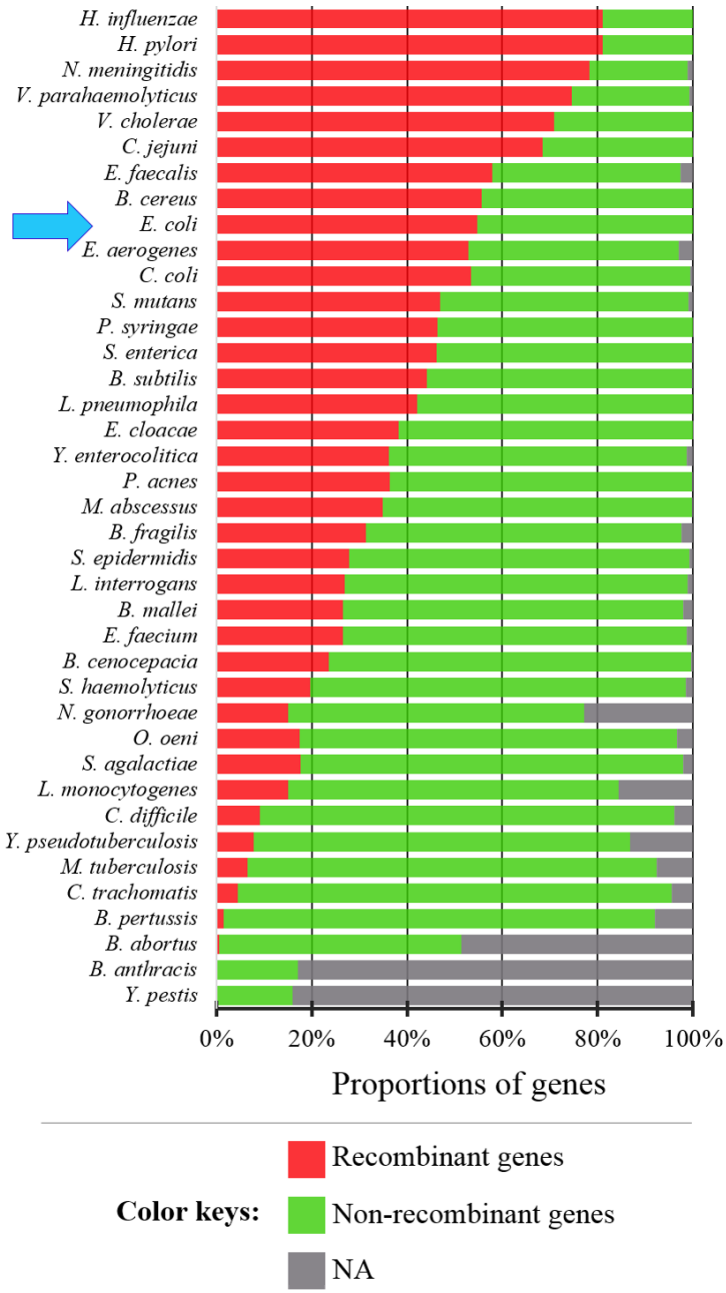

Figure 42. Proportions of recombined core genes in 39 bacterial species. Cases of "NA" (not able to determine the likelihood of recombination) were due to the lack of polymorphism in the gene sequences. Position of *E. coli* was marked by a blue arrow.

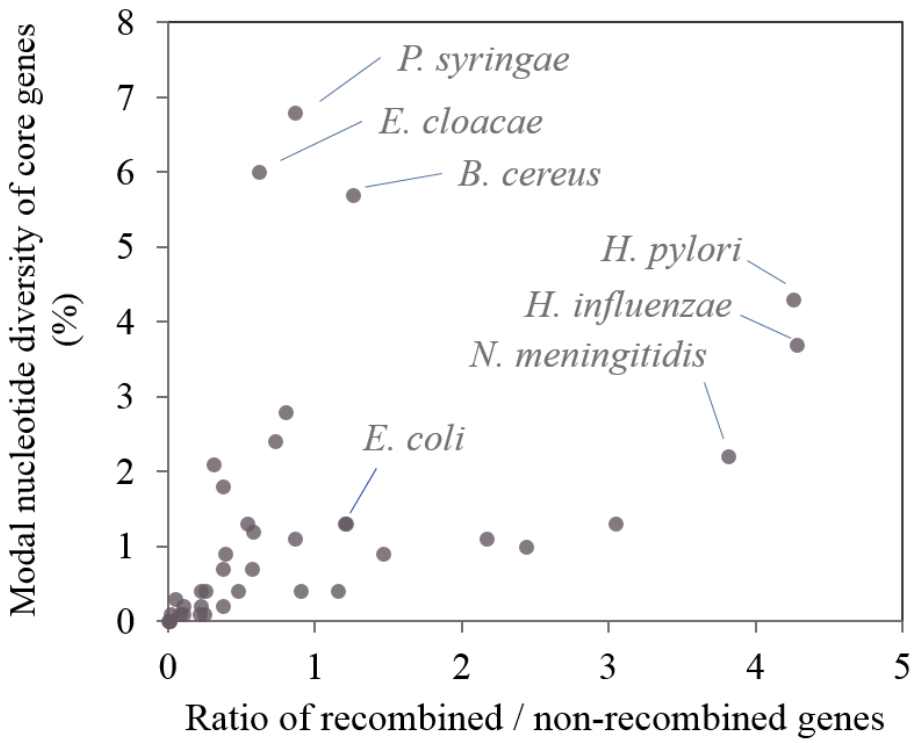

Figure 43. Relationship between the prevalence of recombination in the core-genome and the genetic diversity of species. *P. syringae*, *E. cloacae* and *B. cereus* appeared as apparent outliers from the positive correlation. Note that these species consisted of two or more distantly related subspecies.

3.3.2. Impact of gene gain and loss in the genomic evolution of *E. coli* and the origins of recently gained genes

In chapter 2, the pan-genome of *E. coli* was shown to contain a large number of singleton genes that does not have orthologous counterpart in *E. coli* genomes. The number of singleton genes was 28,007. Two possible mechanisms occurred as the origin of the singleton genes. The genes could be either recently generated paralogs, originating from duplication of another *E. coli* gene, or recently gained genes that were transferred from species other than *E. coli.* Identification of the origin of those genes would comprise an interesting exploration. In order to search for the origins of those singleton genes, amino acid sequences of the genes were searched against all available prokaryotic genomes using ublast local alignment tool. Among 28,007 singletons 12,026 had no significant in any genome. These genes should be considered as ORFans by definition. ORFan gene is defined as a protein-coding gene encoded in a genome that does not have any known homolog in the other genomes. Presence of ORFans might be due to spurious erroneous annotation of protein-coding sequence, or due to the fact that vast majority of prokaryotic gene space has not been sequenced yet. For another 2,603 singleton genes, significant hit was found only in the *E. coli* genomes. Profiles of the number of intra-*E. coli* hits and intra-*E. coli* sequence identity shown in the **Fig. 44** demonstrated that the homologs of these genes had high sequence similarity and present in small number of strains. Based on high sequence similarity, these genes were most likely originated from recent duplications that happened within *E. coli* genomes. Another 2,907 singleton genes had significant matches in the genomes of other species but not in *E.*

coli genome. These genes were highly likely introduced to *E. coli* by recent HGT events. The remaining 10,471 singleton genes had significant matches both in the genomes of *E. coli* and other species. To figure out the route by which such genes came to exist in the *E. coli*, phylogenetic analyses is required. For the 2,907 genes that were highly likely introduced by HGT, taxonomic composition of the best hits were inspected in detail. The taxonomic origin of the best hit proteins of 2,907 putatively xenologous genes spanned 2 domains, 5 phyla, 14 classes, 35 orders, 63 families, 145 genera and 362 species. As shown **Fig. 45** the sequence identity between the *E. coli* gene and its best hit was very high for most cases. Taxonomic composition of the best hits at order or higher level was summarized in the **Table 7.** List of genera which contributed most to the best hit of putative horizontally transferred genes were summarized in **Table 8**. Top ranked genera and species most frequently belonged to the family *Enterobacteriaceae*. Nonetheless many distantly related taxa were also discovered. The species which contained the best hit protein cannot be safely assumed to be the source of gene transfer. However it would be safe to conclude that the species and *E. coli* somehow shared the common environmental gene pool, in which gene transfers can occur between the microbes.

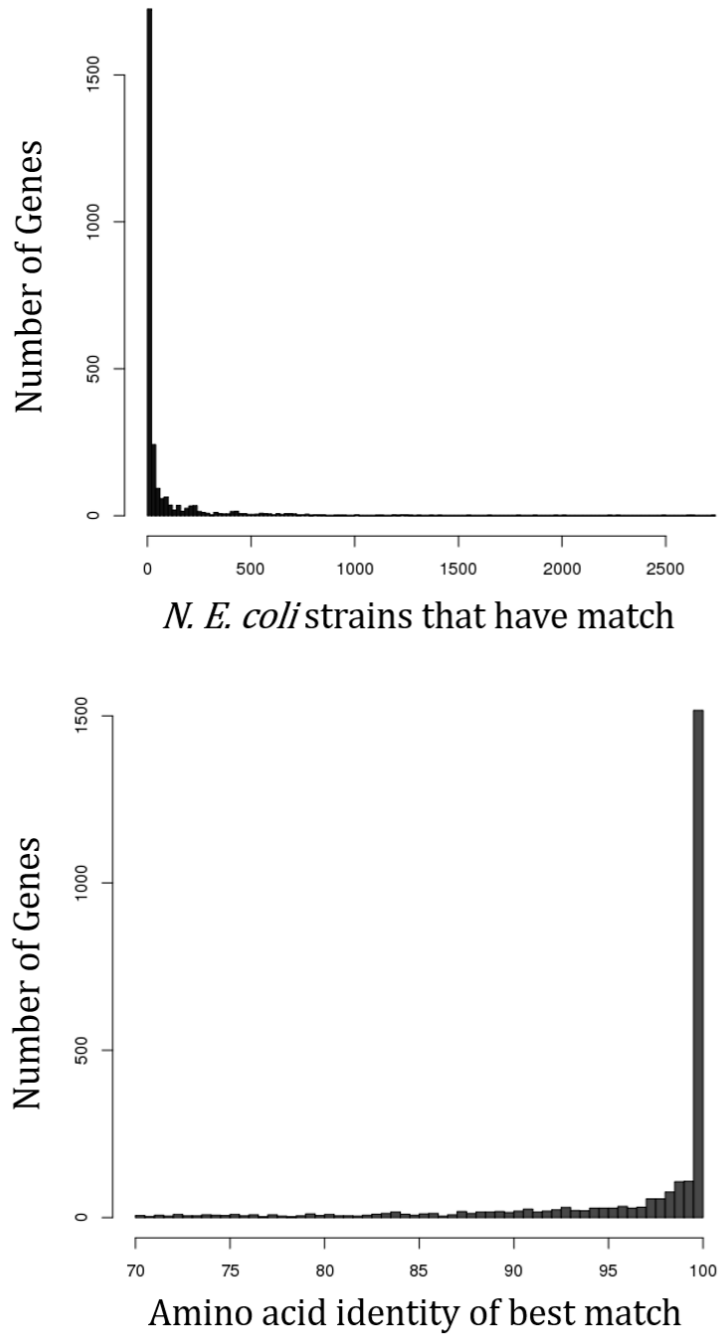

Figure 44. Profiles of the ublast hits of the singleton genes of *E. coli* found in the genomes of *E. coli*..

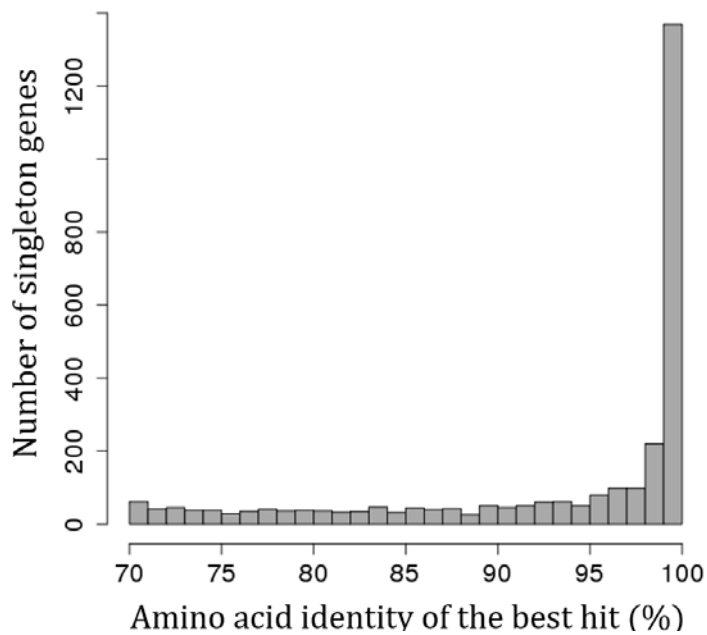

Figure 45. Amino acid identity distribution between the *E. coli* singleton genes and their hits found in the genomes of other species.

Table 7. Taxonomic composition of the best hits of *E. coli* singleton genes whose homologues were found only outside of *E. coli.*

Phylum	N	Class	N	Order	N
Proteobacteria	2578	*Alphaproteobacteria*	80	*Rhizobiales*	79
				Rhodobacterales	1
		Betaproteobacteria	43	*Burkholderiales*	24
				Neisseriales	1
				Nitrosomonadales	1
				Zoogloea_o	17
		Deltaproteobacteria	3	*Desulfovibrionales*	1
				Desulfuromonadales	2
		Epsilonproteobacteria	34	*Campylobacterales*	34
		Gammaproteobacteria	2,418	*Aeromonadales*	14
				Alteromonadales	20
				Chromatiales	1
				Enterobacteriales	2,166
				Methylococcales	4
				Oceanospirillales	6
				Orbales	1
				Pasteurellales	4
				Pseudomonadales	114
				SAR86	2
				Thiotrichales	8
				Vibrionales	58
				Xanthomonadales	20

Table 7. Continued.

Phylum	N	Class	N	Order	N
Euryarchaeota	1	*Methanomicrobia*	1	*Methanosarcinales*	1
Actinobacteria	20	Actinobacteria_c	20	*Bifidobacteriales*	1
				Corynebacteriales	11
				Micrococcales	6
				Propionibacteriales	2
Bacteroidetes	52	*Bacteroidia*	17	*Bacteroidales*	17
		Cytophagia	1	*Cytophagales*	1
		Flavobacteria	19	*Flavobacteriales*	19
		Sphingobacteriia	15	*Sphingobacteriales*	15
Firmicutes	256	*Bacilli*	190	*Bacillales*	103
				Lactobacillales	87
		Clostridia	65	*Clostridiales*	65
		Negativicutes	1	*Selenomonadales*	1

Table 8. The top 25 genera that contributed most to the best hits to the *E. coli* singleton genes.

Top 25 genera	Cases
Klebsiella	576
Enterobacter	538
Salmonella	219
Citrobacter	195
Kosakonia	152
Pseudomonas	65
Methylobacterium	64
Yersinia	64
Clostridium_g4	62
Vibrio	54
Escherichia	47
Acinetobacter	46
Streptococcus	41
Pantoea	39
Serratia	37
Listeria	36
Enterococcus	35
Staphylococcus	34
Campylobacter	34
Bacillus	30
Cronobacter	27
Providencia	27
Raoultella	26
Erwinia	25
Dickeya	24

Majority of the listed genera belonged to *Enterobacteriaceae.* Among the genera included in this top 25 list, those belong to distant classes or phyla were *Methylobacterium* (*Alphaproteobacteria*), *Campylobacter* (*Epsilonproteobacteria), Clostridium, Streptococcus, Listeria, Enterococcus, Staphylococcus* and *Bacillus* (*Firmicutes*).

From the input information of presence/absence matrix of pan-genome genes and the phylogenetic tree of the strains, gene gain/loss events were mapped to the branches of phylogenetic tree. Counts of the gene gain and loss events were divided by branch length to generate relative rate of gene gain and gene loss per core-genome substitution. To investigate the historic trend of gain/loss rates, gain/substitution and loss/substitution rates of the branches were compared with the height of the branches. In the result shown in **Fig. 46**, the internal branches closer to the root showed relatively lower rates of gene gain and loss per branch length and the branches closer to the terminal nodes (the current strains) exhibited higher rates of gene gain and loss per branch length. The ratio between gene gain and gene loss were not even but biased toward greater gene gain rates than loss rates. Inflation of gene gain and loss rates per branch length in the modern branches, along with temporal decline of R/theta, raised a question about the reason behind such trend.

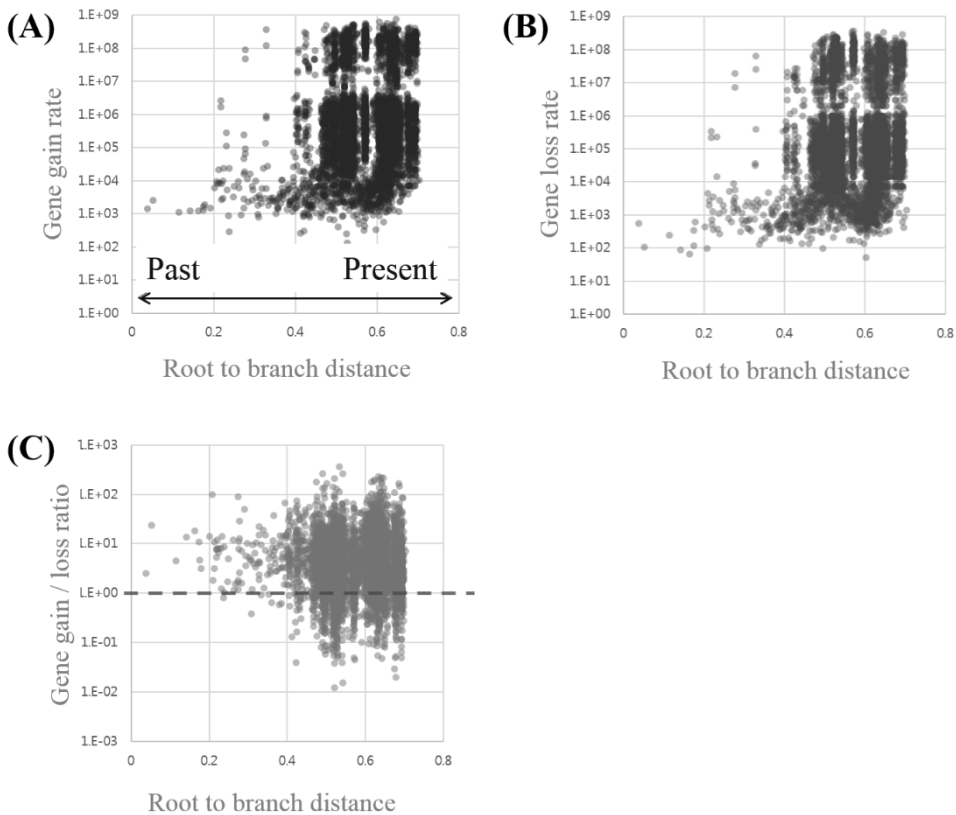

Figure 46. Historical trend of gene gain and loss rates. (A) Gene gain rates. (B) Gene loss rates. (C) Relative ratio of gain/loss rates. Branch-specific estimation of gene gain rates and loss rates were estimated by GLOOME.

3.3.3. Analysis of the signs of natural selection in the pan-genome of *E. coli*

Imbalance between the rates of synonymous nucleotide substitution and nonsynonymous nucleotide substitution has been widely employed as indicator of natural selection. Excessive nonsynonymous substitutions is usually interpreted as the sign of positive (diversifying) natural selection, and excessive synonymous substitutions as the sign of negative (purifying) natural selection. The rates of synonymous and nonsynonymous substitutions (dN and dS) were analyzed for the genes of *E. coli* pan-genome. Genes belonging to different gene frequency categories (5 categories from core genes to rare genes) displayed apparently different distributions of dS but the distributions of dN were similarly centered at 0 in all categories of genes (**Fig. 48**). Based on the assumption that synonymous substitution rates depended on the evolutionary time, the observed trend that core genes exhibited the highest dS could be explained by the fact that the core genes have been evolving in the *E. coli* population for the longest time. Despite of their high dS, the core genes exhibited similar dN with the other categories of genes. It implied that dN was strongly suppressed in the core genes. Ratio of dN/dS estimated for each gene would provide gene-by-gene determination of the direction of selective pressure, so distribution of dN/dS ratio of the genes that belonged to different gene frequency categories were compared. Result shown in the **Fig. 47** clearly indicated that genes with high frequency have experienced strong negative selection and low-frequency genes have experienced relatively relaxed negative selection. Beside the relaxation of negative selective pressure on phyletically rare genes, the results provided a couple of insights. First, negative selection was the predominant direction of natural selection across entire pan-genome. Second, nonetheless there were some genes that were under positive selection.

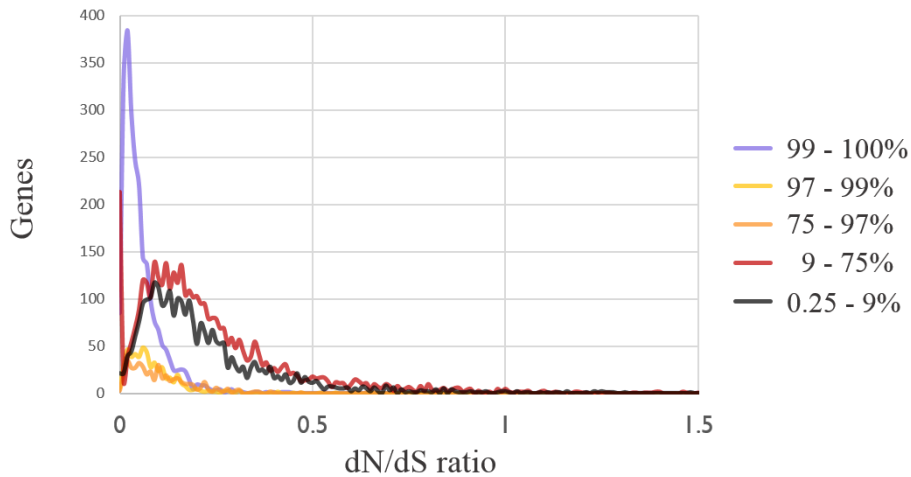

Figure 47. Distribution of dN/dS ratio of the genes in different gene frequency categories.

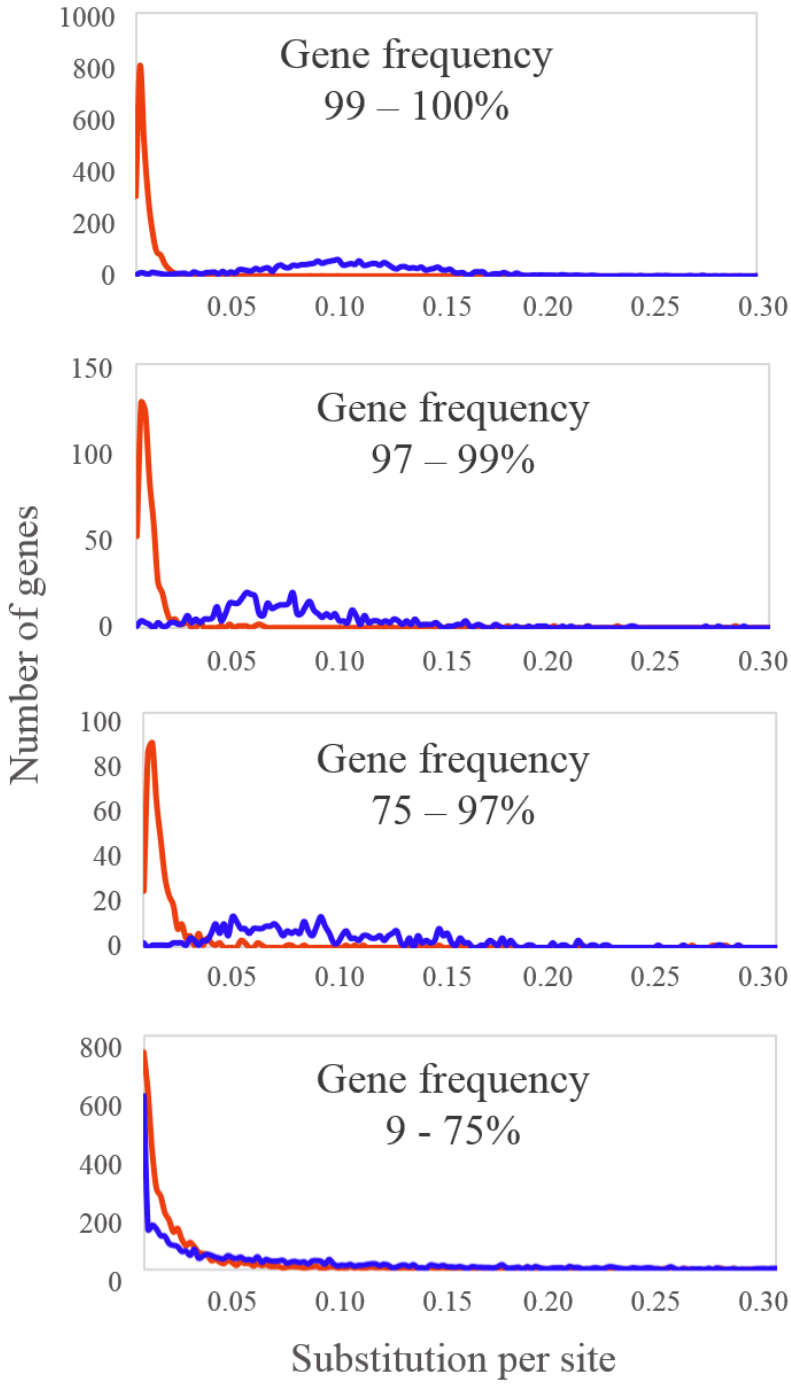

Figure 48. Distributions of dN and dS in the genes belonging to 4 different gene frequency categories. Distribution of dN was depicted by red line and dS by blue line.

Positive selection happens when mutational change of protein sequences confer fitness advantage to the organisms. Collectively, 97 genes were shown to have dN/dS ratio greater than 1. The functions of 74 genes out of the 97 detected positively selected genes could not be identified by KEGG, SEED and eggNOG databases. Description of the functions identified for the remaining 23 positively selected genes were summarized in the **Table 9**. Functions related to DNA mobilization were enriched in the list. Eleven genes were encoded for transposase, retrotransposon protein, or plasmid conjugative transfer protein. Three genes had functions related to invasion of host cells. A gene for drug efflux pump and a gene that regulates the transcription of drug-resistance genes were also included in the list. Functions other than these included phospholipid metabolism, RNA chaperone protein, endopeptidase and Ankyrin repeat protein. Among the core genes that had dN/dS ratio lower than 1, genes under relaxed negative selection were selected based on the cutoff dN/dS > 0.3. The functions of the 23 genes were listed in the **Table 10**. These genes are relatively free from structural constraints, or maybe under positive selection because intra-specific estimation of dN/dS often produces the ratio lower than 1 in the presence of positive selection. Notable functions were included in the list, such as, a flagella biosynthesis protein and an O-antigen biosynthesis gene, 3 genes related to the respiration (nitroreductase, ubiquinone biosynthesis, and anaerobic respiration using trimethylamine-oxide), 2 genes related with uptake of specific substrates (enterobactin for iron uptake, glucose-specific phosphotransferase), 2 genes for reuse of nucleotides (DNA as carbon source, hydrolysis of pyrimidine), 2 genes related with oxidative stress response (an inhibitor of *RpoS* proteolysis, a *soxRS* regulon) and a beta-lactamase. The functions listed here may reflect what were the most important environmental stresses faced by *E. coli* in their natural lifestyle.

Table 9. Functions of the positively selected genes in the *E. coli* pan-genome.

Gene index	Gene frequency	dN/dS	Product	eggNOG category
5987	0.164	1.144	Efflux pump *Lde*	G
20021	0.041	1.674	*TetR* family transcriptional regulator	K
10109	0.130	1.010	Integrase / putative transposase *TniA*	L
8473	0.068	1.089	Retrotransposon protein	L
11432	0.027	1.750	Retrotransposon protein	L
8749	0.212	1.265	Transposase	L
10114	0.110	1.215	Transposase	L
7164	0.075	1.043	Transposase	L
6888	0.048	1.650	Transposase	L
19343	0.027	2.554	Transposase	L
23593	0.027	1.020	Transposase	L
1812	0.986	1.671	Flagellar biosynthesis protein *FliC*	N
16123	0.048	1.711	*P*lasmid conjugative transfer protein *PilJ*	**NA**
18802	0.027	1.210	Phosphatidate cytidylyltransferase	**NA**
6862	0.151	1.135	Transposase *OrfAB*	**NA**
37236	0.048	1.041	*IpaB*/*EvcA* family	O
3623	0.692	2.156	A protein that affects formate dehydrogenase	S
10752	0.288	1.034	Ankyrin Repeat containing protein	S
18440	0.062	1.091	Endopeptidase	S
11481	0.034	1.063	Internalin	S
11482	0.034	1.387	Internalin	S
30801	0.027	1.114	*ProQ*/*FINO* family	S
7186	0.027	1.341	Transfer protein	S

Among 97 genes that had dN/dS > 1, this table included 23 genes. For the 74 genes excluded from this table, it was unable to identify the functions.

Table 10. Functions of the high-frequency genes (present in >97% of strains) that were suspected to be under positive selection or relaxed negative selection.

Gene index	dN/dS	Product	eggNOG category
1812	1.671	Flagellar biosynthesis *FliC*	N
1651	0.412	5-methyl-dCTP pyrophosphohydrolase	L
197	0.453	Methyltransferase *UbiE* in Ubiquinone/menaquinone biosynthesis	Q
782	0.381	Nitroreductase	C
3388	0.418	UDP-glucose:(glucosyl)lipopolysaccharide alpha-1,2-glucosyltransferase	M
3871	0.309	Beta-lactamase	V
2046	0.528	LSU ribosomal protein L25p	J
540	0.443	Putative cytoplasmic protein *YbdZ* in enterobactin biosynthesis operon	S
930	0.311	Chaperone protein *TorD* involved in the biogenesis of *TorA*	S
3845	0.309	Conserved protein	S
3969	0.313	Membrane protein / glucose-specific IIA component in PTS system	S
1029	0.399	Lipoprotein	S
3171	0.316	Membrane protein	S
4030	0.301	Inhibitor of *RpoS* proteolysis, stabilizing sigma stress factor *RpoS* during oxidative stress	K
1810	0.470	Flagella biosynthesis protein *FliZ*	K
1097	0.393	Prophage protein	S
2356	0.313	Cellular component movement	S
781	0.357	Member of the *soxRS* regulon which responds to oxidative stress	S
3162	0.648	DNA utilization protein *HofO*	S

Strength of natural selection per gene was also analyzed for each phylogenetic group. **Fig. 49** illustrated that dN/dS was generally elevated when the rates were calculated for the genomes that belong to the same phylogenetic group. This implication and meaning of this phenomenon cannot be given here, and have to be discussed later. The difficulty came from the fact that it is not clear if the members of the same phylogenetic groups share the more homogeneous ecological niche when compared to the members of different phylogenetic groups do. If they did share the more homogeneous ecological niche, than the efficiency of natural selection would be higher inside the phylogenetic group. If that was the case, than observed elevation of dN/dS in the group would be expected under the presence of positive selection. Under the presence of negative selection, with higher efficiency of natural selection, the dN/dS values should have shifted down inside the phylogenetic group. However, all such interpretations critically depended on the assumption of the differentiation of ecologically niches between the phylogenetic groups. Therefore, without knowledge of the ecological niche of phylogenetic groups no interpretation could be made at this point.

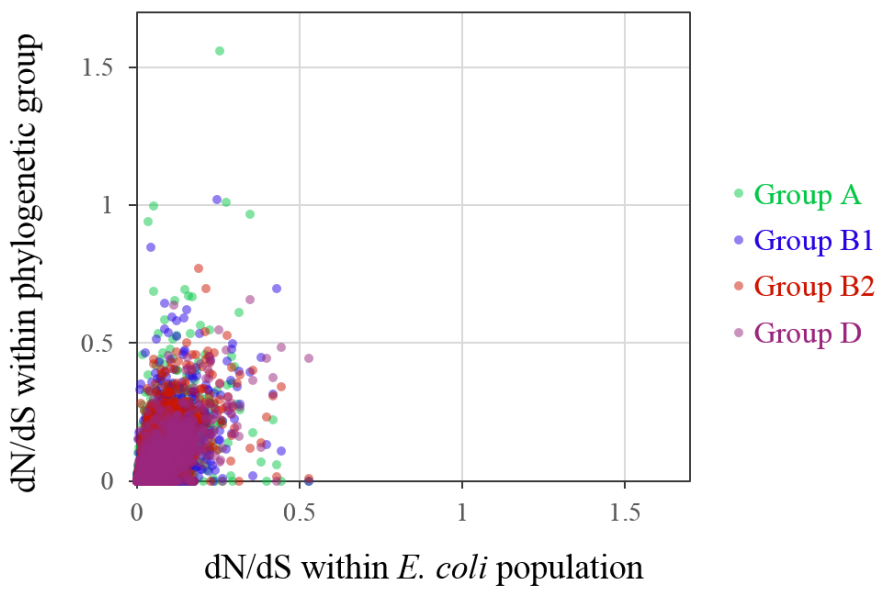

Figure 49. Within group dN/dS estimations vs. species-level dN/dS estimation. When the ratio was calculated within each phylogenetic group, the resulting dN/dS ratio were generally higher than what was estimated using all sequences of *E. coli.*

3.4. Discussion

Evolutionary processes such as mutation, recombination, gene transfer and natural selection work together in the bacterial genome evolution and are influenced by each other. In this chapter the roles played by those evolutionary processes in the course of *E. coli* evolution were identified. In accordance with the observation of increased clonality within the phylogenetic groups of *E. coli*, the relative impact of recombination estimated by R/theta (the base changes introduced by recombination versus the base changes introduced by mutation) decreased in the branches closer to the tip of phylogeny. The distribution of R/theta values estimated for the terminal branches showed the dominant peak at 0.6 – 0.8, meaning that the extant *E. coli* population have R/theta lower than 1. Previous studies of recombination rates of *E. coli* reported the values of 0.32 – 2.14 by MLST analysis (Wirth, et al. 2006), 0.70 by MLST analysis (Vos and Didelot 2008), 0.90 by analysis of 20 genomes (Touchon, et al. 2009), 1.02 by analysis of 27 genomes (Didelot, et al. 2012) and 0.92 by analysis of 19 genomes (Bobay, et al. 2015). Estimation made in this study was slightly lower than the values reported in the previous genome-scale analysis. Lower estimation was possibly due to the fact that the value reported in this study was focused on the "recent" recombination rates. Support for the declining impact of recombination also came from the gene-by-gene detection of recombination events. When the sequences of all strains were used to detect the recombined genes, 54.8% of the core genes were detected to have undergone recombination. However when the sequences of the strains belonging to the same phylogenetic group were used, the proportion of recombination-detected genes dropped significantly. Exceptional cases

were observed for group D and B2 which maintained even higher proportions of recombined genes. Maintenance of high recombination rates in group D and B2 was demonstrated again in the estimation of total length of DNA segments imported by recombination per strain's genome. Recombinational gene flow into the lineages of group D and B2 mostly came from group B2. Higher recombination rates of group D and B2 coincided with the results of the chapter 2. In the analysis of pan-genome growth rate per unit phylogenetic diversity of the strains, group D and B2 showed the least inflated values.

In this thesis, one of the most important hypothesis that the sequence diversification is slower for more clonal populations and faster for more recombining populations. As an circumstantial evidence for that hypothesis, a general trend that the species with higher prevalence of recombined genes have higher sequence diversity of core genes, was observed by comparison of multiple bacterial species. Species with the highest recombination frequency like *H. pylori* and *H. influenza* achieved the largest intra-specific sequence diversity (except for some species that contained multiple sub-species). Species with the least recombination frequency like *B. anthracis* and *Y. pestis* achieved the smallest intra-species sequence diversity. The ages of species might have been the hidden factor behind the pattern. However, the pattern of correlation was so clear that some positive connection must be present between recombination's frequency and the intra-species sequence diversity. In addition theoretically it is apparent that the lack of recombination would produce genome-wide linkage and accumulation of genetic variation in strong genome-wide linkage would be at high risk of being purged by any advantageous mutation.

Observation of open pan-genome led to a question regarding the origin of genes gained by *E. coli* population. Singleton genes in the pan-genome comprised a good model to study the source of genomic influx. For 10% of the singleton genes, the closest matching sequences were found to be another *E. coli* gene, which meant that those singleton genes were produced by duplication of the genes that already had been present in *E. coli*. For another 10% of the singleton genes, it was able to find highly similar homologs in the genomes of other bacterial species. The species that contained the best-matching homolog and *E.* coli likely have shared the common environmental gene pool. Taxonomic profile of best-matching homologs was made up preferentially of the other members of *Enterobacteriaceae* family. Nonetheless numerous species that are distantly related to *E. coli* were found too, especially in the phylum *Firmicutes* (*Clostridium, Streptococcus, Listeria, Enterococcus, Bacillus, Staphylococcus*). Disproportionate enrichment of taxonomically close relatives might be a hint that overall genomic sequence similarity has a positive impact on the likelihood of gene transfer. Previous studies actually have demonstrated that HGT is biased, and more likely to occur between close relatives (Skippington and Ragan 2012; Williams, et al. 2012). Phylogenetic proximity is a factor that determine the likelihood of HGT events, as well as the overlap of ecological niche (Andam and Gogarten 2011). Based on that notion, the observed high frequency of singleton genes shared with the members of *Enterobacteriaceae* was likely a result of phylogenetic bias in HGT. On the other hand the fact that *Firmicutes* and *Bacteroidetes* were most frequently detected to share singleton genes of *E. coli* at phylum level was likely resulted from ecological overlaps between *E. coli* and the species in those phyla. Interestingly the two phyla were the most abundant phyla in animal gut microbiome. For 40% of the singleton genes it was not able to find

significant matches in the 44,140 available prokaryotic genomes. Those singleton genes could be concluded as ORFans at this moment. Abundance of ORFans in *E. coli* pan-genome implies that the current database of genome sequences can only partially cover the prokaryotic gene diversity of natural microbiomes that can interact with *E. coli*.

Analyses of dN/dS ratio within *E. coli* species revealed two trends, (i) that the genes were dominantly under strong negative selection and just a few genes were under positive selection and (ii) that genes with high gene frequency (core genes) experienced stronger negative selection while genes with low gene frequency (phyletically rare genes) experienced relaxed negative selection. Among the genes found in 97% or more *E. coli* genomes (roughly defined core genes), only two genes showed dN/dS ratio greater than 1. The two genes were a flagellar biosynthesis gene *fliC* that encodes the filament of the flagella in *E. coli* and diverse other bacteria (Liu and Ochman 2007) and a protein coding gene of unknown function. Detection of strong positive selection in *fliC* meant that rapid adaptive changes has been required for the filament protein encoded by this gene. Relaxation of structural constraints (dN/dS ratio < 1 but relatively high) was observed for the genes belonging to diverse functional categories. Notable among them are beta-lactamase, enterobactin biosynthesis gene, and the genes that respond to oxidative stress. Among non-core genes (gene frequency <97%) 95 genes were found to have dN/dS ratio greater than 1. Only 12 out of the 95 positively selected genes were able to be functionally identified using eggNOG, KEGG and SEED databases. Among the functions identified, DNA mobilization-related genes were most abundant. Ratio of dN/dS was higher within each *E. coli* phylogenetic group than among the whole *E. coli* strains.

This trend could be generated if the level of ecological coherence is higher within each phylogenetic group, because in the group of ecologically in-coherent strains the impact of positive selection will be diluted by the strains not affected by the selection. In the group A and B1, a function-unknown membrane protein, a secretion pathway protein and a hemolysin genes were detected to be positively selected.

Presence of strong positive selection is a condition that realize selective sweep in clonal populations. In the analysis of this chapter just a few number of genes were confirmed to be under positive selection. The frequency of positive selection might seem insufficient to support the hypothesis of repression of sequence divergence within the phylogenetic groups by selective sweeps. However, by a number of reasons the possibility of selective sweeps cannot be rejected. First, dN/dS ratio measured between intra-population samples was shown to be much variable under the same selection coefficient than what measured for divergent species because the statistical power is weakened by low dN and dS values (Kryazhimskiy and Plotkin 2008). From intra-population samples, according to a simulation study dN/dS values can be smaller than 1 even if positive selection was given. On the other hand dN/dS values greater than 1 were not generated under negative selection (Kryazhimskiy and Plotkin 2008). Therefore, while the presence of positive selection can be confirmed when $dN/dS > 1$, the absence of positive selection cannot be confirmed only by $dN/dS < 1$. Ideally, as a complementary test for dN/dS which are powerful at long time scale, positive selection within short time scale can be detected by long-range haplotype test for highly co-linear genomes (Shapiro, et al. 2009; Vitti, et al. 2013). Unfortunately *E. coli* core-genome is intercalated by numerous strain-specific segments so that long-range haplotype test would not be highly efficient.

Second reason is that selective sweep was recently shown to be driven also by negative frequency-dependent selection (Takeuchi, et al. 2015). Third, relatively large number of dispensable genes were detected to have been under positive selection. Whether or not the selection acting on dispensable genes can drive the genome-wide selective sweep is unknown. In theory, if fitness advantage was gained by a low-frequency gene, selective sweep in the core-genome would happen within the ecologically coherent group with the rise of the gene frequency of that dispensable gene. Observed low diversity of pathogenic clades (SII, SIII and E) and known ecological importance of laterally acquired genes in those clades may reflect the cases of strong positive selection acted on the dispensable genes. What intended to be emphasized in this section was that even if the sign of positive selection was scarce in the genomes of *E. coli*, possibility of selective sweep is not rejected. Regardless of the mechanism that drive it, recent selective sweep events are expected to leave characteristic local depletion of population level polymorphism. Therefore, to not rely on ambiguous assumption of the possibility of selective sweeps, further efforts should be made to look for the direct signs of selective sweeps.

CHAPTER 4

Systematics study of *E. coli* and related taxa

4.1. Introduction

4.1.1. Timed history of bacterial evolution

Based on the observed clocklike manner of the accumulation of mutations in 16S rRNA in multiple diverse bacterial lineages, Ochman and Wilson committed molecular dating of a number of important divergence points in bacterial evolution (Ochman and Wilson 1987). In order to calibrate the molecular time-scale to the geological time-scale, calibration points whose dates could be refined by external information were necessary. For bacteria, fossil evidences that can be used to allocate the ancestral organisms to geological time zone are very limited. However, genius minds of these researchers were able to extract multiple calibration points within the history of bacterial evolution: (a) Cyanobacterial fossils provided a calibration point when the clade containing modern cyanobacteria is at least 1.3 billion years old. (b) Based on the geological estimates for the atmospheric oxygen level, the minimal level of atmospheric oxygen required for existence of aerobic organisms was realized not longer than 600-800 million years ago. (c) Obligate symbiosis of chloroplast, mitochondria, nodule bacteria and luminous bacteria with the macro-organisms provided the logical bases to apply the fossil ages of host macro-organisms to the ancestor of modern symbiotic bacteria. Ochman and Wilson estimated that the divergence between *E. coli* and *S. enterica* serovar Typhimurium took place in 120 – 160 million years ago, and the divergence between *E. coli* and *Shigella* 20 – 30 million years ago (Ochman and Wilson 1987). Both of the estimates were supported by circumstantial justifications. The estimated divergence time of *E. coli*

and *S. enterica* serovar Typhimurium was coincided with the origin of mammals. The ability to ferment lactose is characteristic to E. coli and lactose is considered to be originated from mammals. The estimated 20 – 30 MYA divergence of *E. coli* and *Shigella* coincided with the estimated age of higher primates. The fact that *Shigella* is specifically associated with higher primates and not the other mammals was considered as a justification for the estimated timing. Later in 2004, a study of Battistuzzi reexamined the multiple divergence timings of bacteria based on 32 globally conserved protein-coding genes (Battistuzzi, et al. 2004). The study applied calibration points that are all related with the origin of the *Cyanobacteria*: the Great Oxidation Event at 2.3 BYA as a minimum bound, the divergence between the *Cyanobacteria* and the other bacteria at 2.04 – 3.08 BYA based on another study, and the appearance of 2α-methylhopanes at 2.7 BYA as the minimum constraint for the age of the *Cyanobacteria*. The reported divergence time between *E. coli* and *S. enterica* serovar Typhimurium was 102 MYA with confidence interval of 57 - 176 MYA (Battistuzzi, et al. 2004). This estimate was consistent with that of Ochman and Wilson. Until now the estimates from the above-mentioned studies are used as the reference calibration dates in more recent studies.

Now with much more comprehensive data available, it was able to refine the age of *E. coli* and its subgroups using the estimates from the above studies as calibration point. The motivations behind this work were, (a) divergence between *E. coli* and *Salmonella* no longer represent the age of *E. coli* because several species that are more closely related to *E. coli* are known, and (b) timing of the divergence between *Shigella* and *E. coli* is no longer valid. *Shigella* and *E. coli* are not sister taxa to each other from current view, as the study of Pupo (Pupo, et al. 2000) and

the following series of studies revealed that *Shigella* are composed of multiple lineages that evolved from diverse branches of *E. coli* (so *Shigella* is a polyphyletic group that share the characteristic phenotypes by convergent evolution). Instead, it seems more relevant to estimate the divergence between the phylogenetic groups of *E. coli*. In the first part of this chapter the evolutionary relationships within the family *Enterobacteriaceae* was analyzed and the ages of *E. coli* species and its subgroups were determined using the core genes of *Enterobacteriaceae*.

4.1.2. Obscurities in the systematics of *E. coli*

The first member and the type species of the genus *Escherichia* was *E. coli*. The second member of the genus, *E. fergusonii* was described in 1985 (Farmer, et al. 1985) and the third member, *E. albertii* was described in 2003 (Huys, et al. 2003). With the 3 known species, the genus seemed to harbor low level of diversity. However the presence of five "cryptic lineages" of the genus *Escherichia* was recognized in 2009 (Walk, et al. 2009). The cryptic lineages (clade I – V) have not been classified as novel species until recently. In 2015, the strains of clade V were proposed as a novel species *E. marmotae* (Liu, et al. 2015). Consequently there are now 4 named species and 4 unclassified clades in the genus *Escherichia*. Two species are misclassified as *Escherichia*, *E. hermannii* (Brenner, Davis, et al. 1982), *E. vulneris* (Brenner, McWhorter, et al. 1982). It is surprising that one of the best-studied bacterial species *E. coli* is not clearly defined in systematics point of view. One reason to say like that is because the species-level identity is not settled for the strains of clade I and clade IV of *Escherichia*. Should we consider these strains as *E. coli* or as independent species? Another reason is that the relative order of branching between the species in the genus *Escherichia* is not clear up-to-date. Is *E. albertii* a more recent relative to *E. coli* than *E. fergusonii* is, or vice versa? The final reason is that current definition of genus *Shigella* is a phenotype-based definition and poses a lot of complications from systematics point of view. The evolutionary origin of *Shigella* has not been explained with certainty since the previous studies are not consistent in their conclusion for the number of events where *Shigella* strains emerged from *E. coli*. Furthermore none of the 4 known species of *Shigella* are monophyletic. In the second part of this chapter the evolutionary relationships within

the genus *Escherichia* was reviewed in detail and the genomic features that distinguished *E. coli* from its closest neighbors were characterized.

4.2. Materials and methods

4.2.1. Reconstruction of *Enterobacteriaceae* phylogeny

The dataset used for reconstruction of *Enterobacteriaceae* phylogeny consisted of the complete genome sequences of the strains belonging to the family *Enterobacteriaceae*. For species or sub-species that have more than 1 complete genomes, only 1 genome was selected and used. An exception was for *Escherichia*, as 6 genomes of *E. coli* were used to represent phylogenetic group diversity and 3 incomplete genomes were used to represent *Escherichia* clade I, IV and V. All genome sequence assemblies were obtained from NCBI FTP (ftp://ftp.ncbi.nlm.nih.gov/genomes/). In total, 63 genomes from diverse species of *Enterobacteriaceae* were included in the analysis.

The orthologous genes in *Enterobacteriaceae* were clustered by using OrthoMCL algorithm (Li, et al. 2003) as implemented in the GET_HOMOLOGUES package (Contreras-Moreira and Vinuesa 2013). As algorithm parameters, cutoff of minimum alignment coverage of 70% was used to filter the blast result, blast e-value cutoff of 10^{-8} was used, and inparalogs were excluded. In the resulting 8,211 orthologous gene clusters which had 2 or more members, 822 gene clusters were present in all 63 genomes. DNA sequences of the 822 core genes of *Enterobacteriaceae* were aligned by codon-based alignment method described in the chapter 2 and 3, and were used in the subsequent phylogenetic analysis.

The alignments of *Enterobacteriaceae* core genes were concatenated to a single supermatrix. This supermatrix was used for 2 types of phylogenetic analysis. First a phylogenetic network was reconstructed by the Neighbor-Net method provided by SplitsTree4 package (Huson and Bryant 2006). Neighbor-Net is a distance-based method, an extension from Neighbor-Joining phylogenetic method that can represent conflicting phylogenetic signals in the sequence data (Bryant and Moulton 2004). Composite histories of core genome can be visualized by the lateral branches produced in this method. A classical bifurcating phylogenetic tree was also reconstructed using the Neighbor-Joining algorithm of MEGA7.0 (Tamura, et al. 2013), based on the nucleotide substitution model of Tamura and Nei (Tamura and Nei 1993). A bootstrap test was performed by 1,000 replications.

4.2.2. Molecular clock analysis and species tree analysis of *Escherichia*

A single calibration point was used for time-tree analysis, which was the divergence time estimated for *E. coli* and *S. enterica* serovar Typhimurium previously reported (Ochman and Wilson 1987; Battistuzzi, et al. 2004). From the result of phylogenomic analysis of *Enterobacteriaceae*, it was observed that *Escherichia, Salmonella, Citrobacter* and *Enterobacter* together form a clade within the family. Therefore a time-tree analysis was performed with the dataset consisted of the species of *Escherichia, Salmonella, Citrobacter and Enterobacter.* As an outgroup *P. vagans* C9-1 was used. RelTime method implemented in the MEGA7.0 (Tamura et al. 2012) was carried out with the constraints given on the timing of the most recent common ancestor of *E. coli* and *S. enterica* (120 MYA as lower limit and 160 MYA as upper limit). In total, 504,238 nucleotide sites were used in the analysis. Local clock was used to allow variation among the sites. Hasegawa-Kishino-Yano (HKY) model of nucleotide substitution was used (Hasegawa, et al. 1985). Rate variation among sites was assumed to be Gamma-distributed with invariant sites, with 4 discrete gamma categories. Gapped sites were ignored and only the 1st and 2nd codon positions were used in the analysis.

The Bayesian multispecies coalescent tool *BEAST (Heled and Drummond 2010) was used to reconstruct species tree. The method can infers the species tree by coalescent method when dataset contains multiple individuals (strains) per each species. For multi-species coalescent analysis randomly selected subsets of 60 genes from the 822 core genes were used. Three different subset of core genes were used

to minimize the effect of gene selection. Strains used in the analysis were 34 strains that belong to *Escherichia*. For each phylogenetic group of *E. coli* and for *E. albertii*, *E. fergusonii* and *E. marmotae*, and for clade I and IV, 3 or 4 strains were included in the data set. In the *BEAST parameter setting, site model was set to have gamma distribution with invariable sites, with 4 different gamma categories and 0.5 proportion of invariable sites. Substitution model of Hasegawa-Kishino-Yano was used with *Kappa* = 2.0. Nucleotide frequencies were set to be empirically estimated. The clock model was set to be relaxed clock with log normal distribution. Population size change was assumed to have followed linear function. Ploidy of each gene was assumed to be Y chromosome or mitochondrial. Coalescent model of Bayesian Skyline model was selected. MCMC chain was set to have 5,000,000 generations with 1,000,000 (20%) burn-in generations. The resulting species trees were summarized by the treeannotator program included in the BEAST 2 package (Bouckaert, et al. 2014) and the resulting maximum-credibility tree was visualized by the FigTree v1.4.2. The species trees were also visualized by DensiTree (Bouckaert 2010) to express the whole information including conflicts contained in the set of species trees.

4.2.3. Reconstruction of *Shigella* virulence plasmid phylogeny

The complete genomes of all 4 known *Shigella* species contained a large plasmid that are around 200 Kb and contain 220 – 270 protein-coding genes. For each of 4 *Shigella* species, one plasmid was selected from the strains that have been regarded as model strains. In order to systematically determine the strains that carry this virulence plasmid (VP), the pan-genome matrix was analyzed. For each of the 4 reference VPs, a presence/absence matrix of the orthologous counterparts of the genes encoded in the 4 reference VPs was compiled based on the pan-genome matrix described in the chapter 2. From the resulting gene presence/absence matrix that spanned 3,909 *E. coli*/*Shigella* strains and 220-270 genes encoded in each plasmid, strains that carry the plasmid and the strains that do not carry the plasmid were delineated. Then, for the 62 genomes that were determined to contain VPs, the plasmid genes that occurred in 50 or more strains were selected as the core plasmid genes. The 83 core plasmid gene sequences in the 62 strains were aligned, concatenated and used to reconstruct the phylogeny of *Shigella* VPs. Codon-based alignment was performed as described in the previous section, and the phylogenetic reconstruction was performed by FastTree following the same method used in the whole-genome phylogeny reconstruction discussed in the chapter 2. To compare the plasmid phylogeny against chromosomal phylogeny, a phylogenetic tree was reconstructed using the core-SNP data. The chromosomal phylogeny contained the strains that lacked plasmid genes. Instead of using all genomes, a set of strains was selected based on the clustering of SNP differences. In addition to the 62 strains that were used in plasmid phylogeny analysis, 290 plasmid-undetected strains were selected for chromosomal phylogeny analysis. SNP sites and alleles were discovered

by the protocol described in chapter 2. After filtering for minimum call rate 0.01 and minimum minor allele frequency 0.01, the alleles at the remaining 195,017 sites were used for maximum-likelihood phylogenetic tree reconstruction.

4.2.4. Reconstruction of *rut* and *phn* operon phylogenies

E. coli-specific core genes were defined using the gene presence/absence matrix of the pan-genome reconstructed in the chapter 2. The genes that were present in >90% of the strains of *E. coli*, *Shigella* and clade I and at the same time absent in >90% of the strains of *E. fergusonii*, *E. albertii*, *E. marmotae*, clade IV and clade V were assigned as *E. coli*-specific genes. Because two operons were included in the discovered *E. coli*-specific genes, multigene phylogenetic trees were reconstructed for the *rut* operon and *phn* operon. First, the homologs of this operon in other prokaryotic genomes were searched. Ublast was run with query length coverage cutoff 0.8, identity cutoff 0.8 and e-value cutoff 10^{-15} against proteins encoded in the 44,140 prokaryotic genomes. Results of ublast search were parsed to make a homolog presence matrix for the strains in the genome database. For *rut* operon, 431 strains were detected to have homologs of all 5 *E. coli*-specific *rut* genes. For *phn* operon, 471 strains were detected to have homologs of all 13 *E. coli*-specific *phn* genes. Sequences of the homologous protein-coding genes found in these "core taxa" were collected and aligned by the codon-based alignment protocol that was described multiple times in the previous sections. Concatenated matrix of *rut* operon homologs and *phn* operon homologs were used for reconstruction of *rut* operon phylogeny and *phn* operon phylogeny, respectively.

4.3. Results

4.3.1. Phylogenomic analysis of evolutionary relationships of *Enterobacteriaceae* species

An overview of evolutionary relationships between the species of *Enterobacteriaceae* was made by phylogenomic analysis of the selected genome sequences. Clusters of orthologous genes were reconstructed from the protein-coding genes contained in the 63 selected genome sequences. Among the clusters of orthologous genes, 822 orthologous clusters were present in all 63 strains and defined as the core-genome of *Enterobacteriaceae*. Two types of phylogenetic analysis was performed on the aligned sequences of the *Enterobacteriaceae* core genome. First, a phylogenetic network was reconstructed by Neighbor-Net algorithm. Resulting phylogenetic network (**Fig. 50**) showed the summarization of the phylogenetic information contained in the 822 core genes. Characteristic features of this phylogenetic network were (i) that terminal branches were long and basal branches were short so that the relationship was close to star-phylogeny, and (ii) that lateral branches are abundant at the basal area. Abundance of lateral branches around the splits between the genera implied that the divergence order of the genera cannot be clearly defined by the core-genome data. The genus *Escherichia* formed a monophyletic clade with *Citrobacter* and *Salmonella*. The members of *Enterobacter* were split so that *E. aerogenes* was affiliated with *Klebsiella* and *Raoultella*. Secondly, a classical phylogenetic tree was reconstructed by Neighbor-Joining method. Even in the presence of conflicting phylogenetic signals (the lateral branches shown in the **Fig. 50**) the NJ tree was fully resolved, with 100% bootstrap

scores obtained for all bipartitions (**Fig. 51**). According to this tree, the closest neighbors of the genus *Escherichia* were *Citrobacter, Salmonella* and *Enterobacter* except for *E. aerogenes.* The four genera were then subjected to subsequent phylogenetic network analysis. The phylogenetic network of 6 species of *Escherichia,* 3 (sub)species of *Salmonella,* 3 species of *Citrobacter* and 4 (sub)species of *Enterobacter* was obtained as **Fig. 52**. Characteristically, the divergence of *Escherichia* spp. was accompanied by significant lateral branches (significant conflicts in the phylogenetic signals). The order of divergences between the species or species-level clades within *Escherichia* remained ambiguous after the phylogenetic network analysis and phylogenetic tree analysis using the 822 conserved genes.

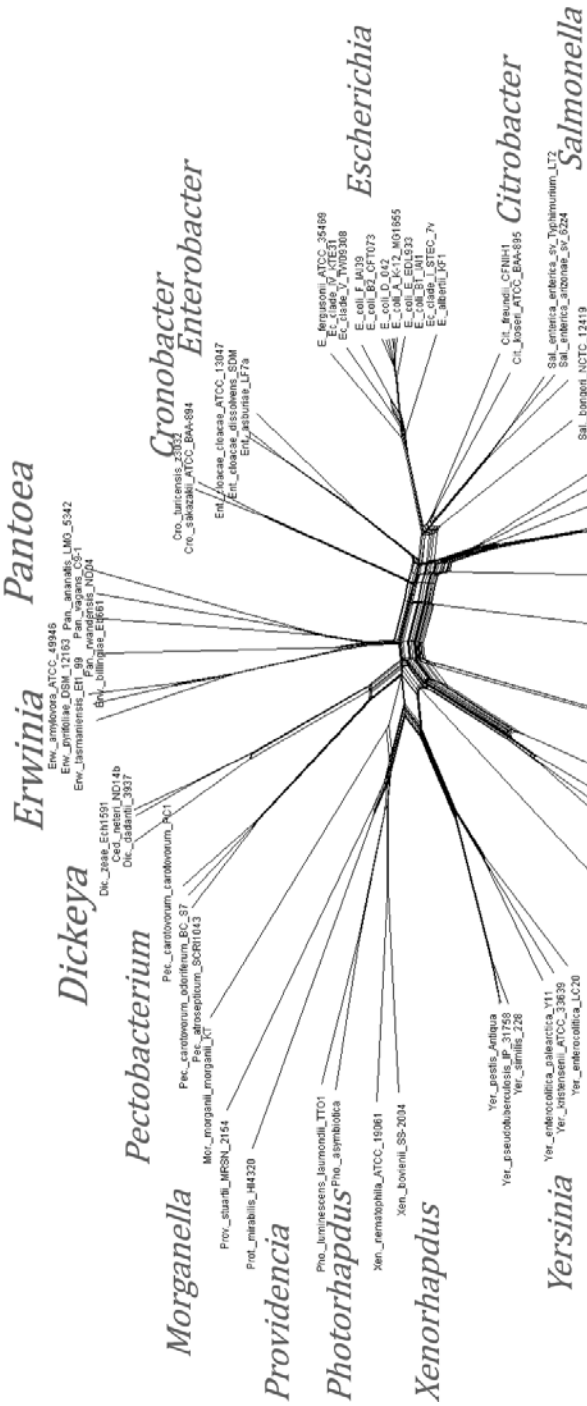

Figure 50. Neighbor-Net phylogenetic network of *Enterobacteriacaeae* based on the core-genome.

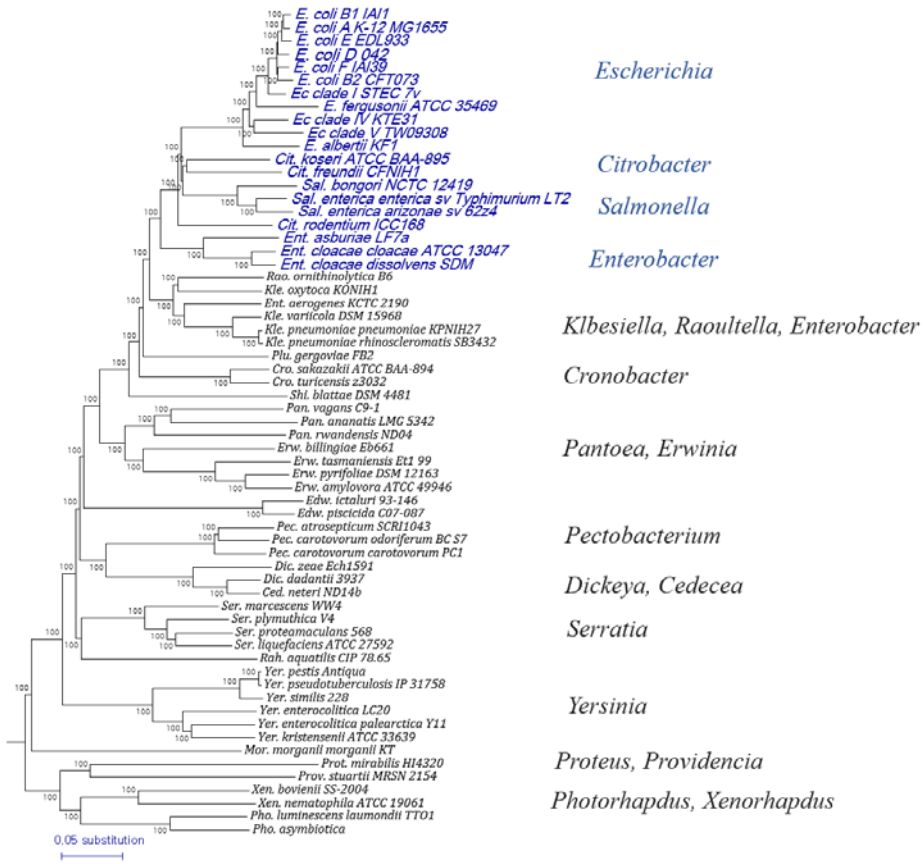

Figure 51. Neighbor-Joining tree of the *Enterobacteriaceae* based on the core-genome. The genus *Escherichia* and its closest genera were highlighted with blue text.

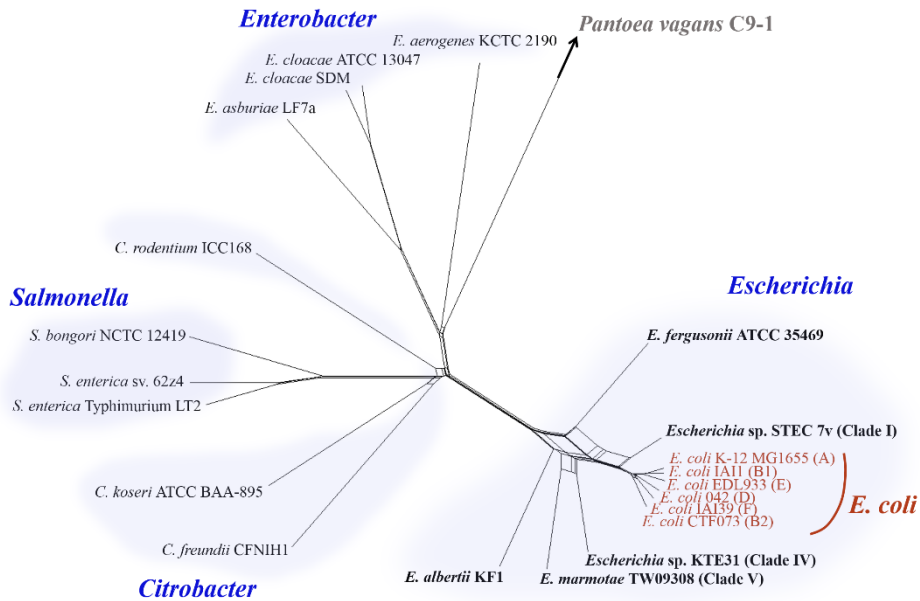

Figure 52. Neighbor-Net phylogenetic network of *Escherichia, Citrobacter, Salmonella* and *Enterobacter* strains.

Branching order within the members of *Escherichia* was further investigated by the multi-species coalescent method, *BEAST. Species-trees were inferred for each of 3 different subsets of *Enterobacteriaceae* core genes, for the species in *Escherichia* and the phylogenetic groups within *E. coli.* The resulting maximum-credibility trees from 3 gene sets were shown in **Fig. 53 (A)-(C).** According to the result, the consensus species trees generated from multi species coalescent of different gene sets agreed on the latest divergence of clade I from *E. coli.* But placement of divergence order of *E. fergusonii*, *E. albertii* and *E. marmotae* was disagreed between the 3 species trees. Since clear resolution was not made by the reduction of species trees to a single maximum credibility tree, the collective information in the coalescent species trees was visualized by DensiTree to look at all information including the conflicts in the data. The resulting DensiTree graphs were shown in **Fig. 54 (A)-(C).** The graphs indicated that coalescent time estimated in different genes varied greatly for the divergence of *E. albertii*, *E. fergusonii*, *E. marmotae* and the clade IV. One consistently supported speculation in the graphs was that the divergence of *E. albertii*, *E. marmotae* and the clade IV from *E. coli* and the clade I happened at the same time and took long time. Clade I seemed likely to be the member of *E. coli* rather than separated species, based on the clarity of the lineage of clade I + *E. coli* and the relatively short time-scale of the divergence of clade I from *E. coli*

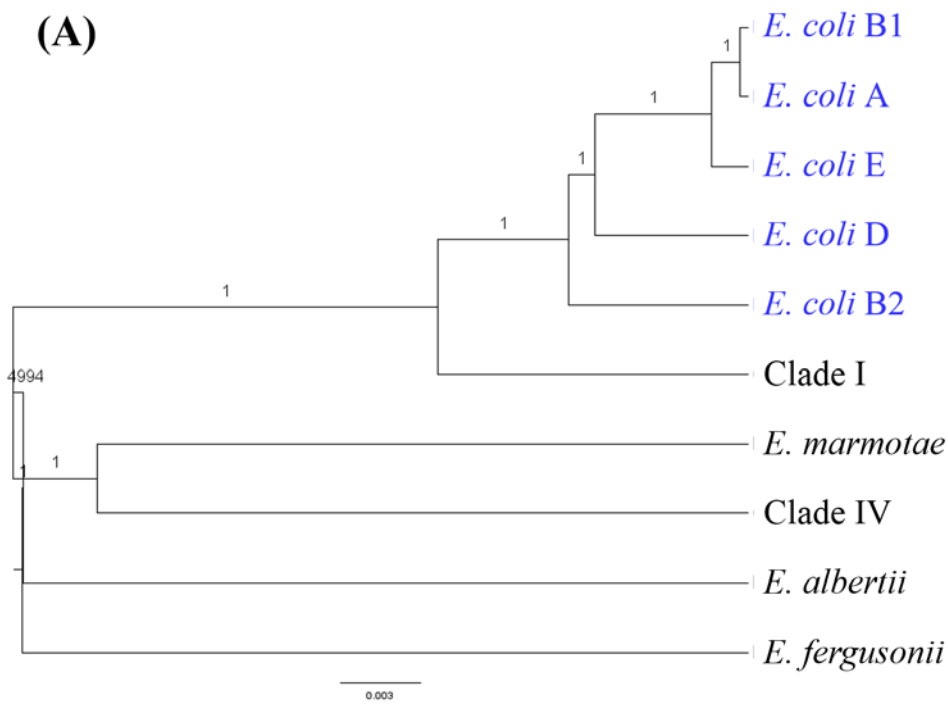

Figure 53. Maximum credibility species tree derived from species trees inferred by multi species coalescence. (A) Result from the first random subset of 60 genes selected from *Enterobacteriaceae* core genes.

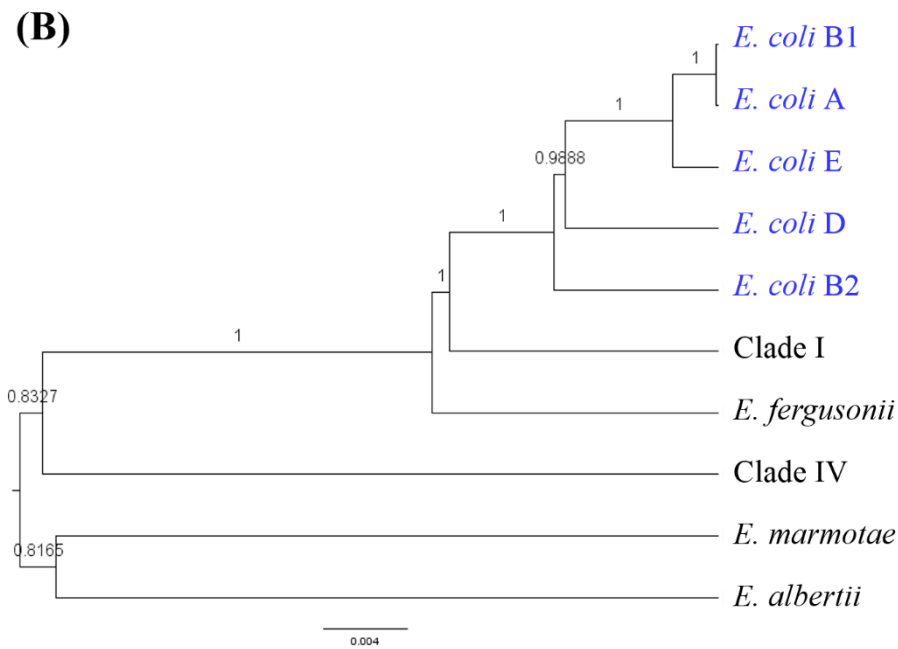

Figure 53. Continued. (B) Result from the second random subset of 60 genes selected from *Enterobacteriaceae* core genes.

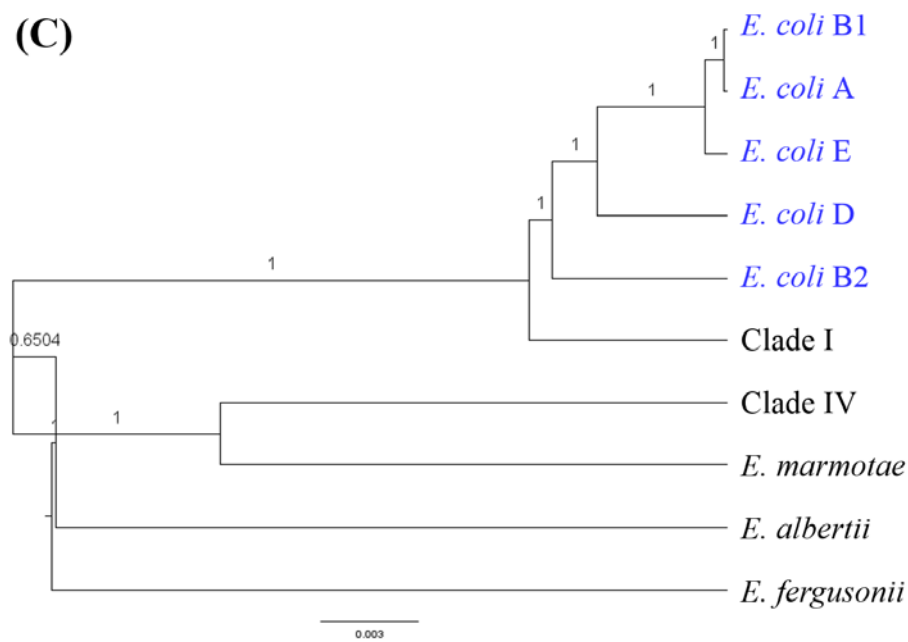

Figure 53. Continued. (C) Result from the third random subset of 60 genes selected from *Enterobacteriaceae* core genes.

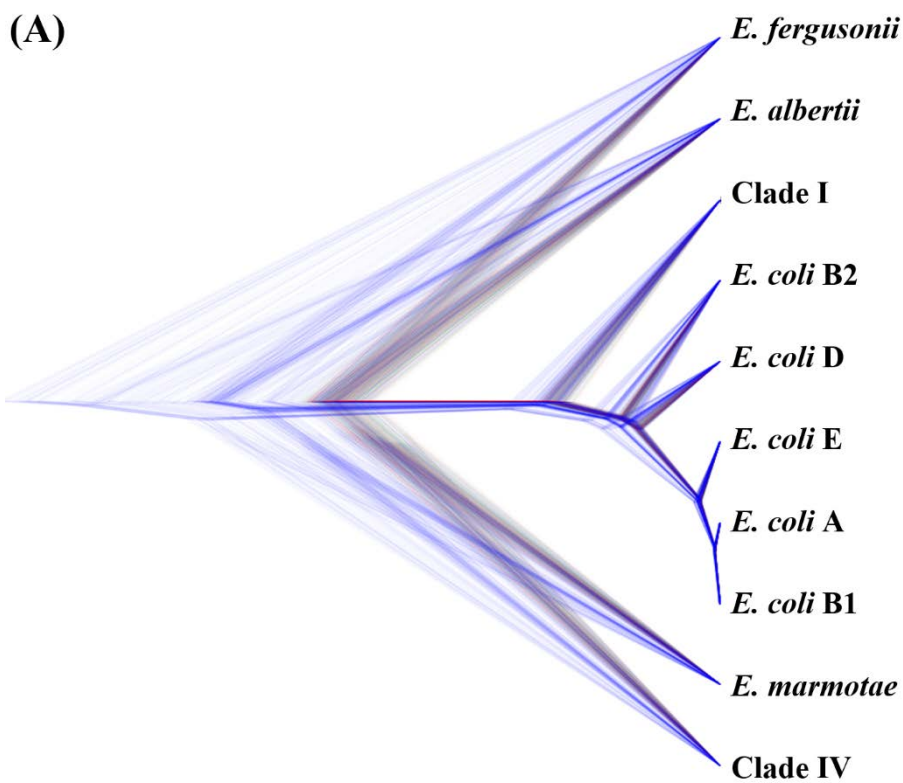

Figure 54. DensiTree graphical representation of the signals generated in multiple species trees. (A) Result from the first random subset of 60 genes randomly selected from *Enterobacteriaceae* core genes.

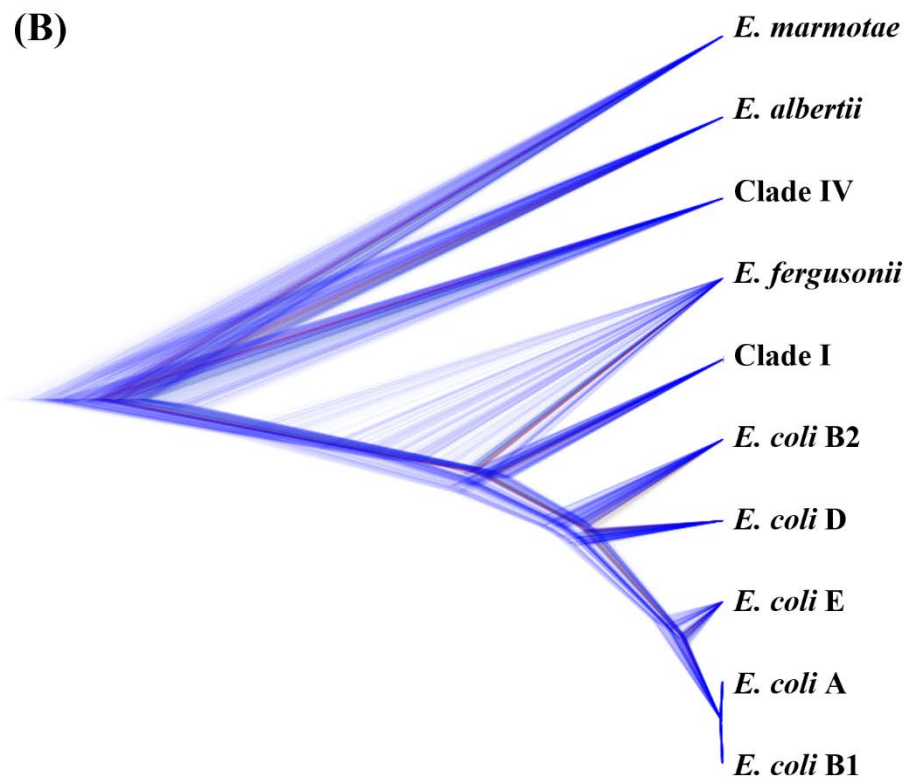

Figure 54. Continued. (B) Result from the first random subset of 60 genes randomly selected from *Enterobacteriaceae* core genes.

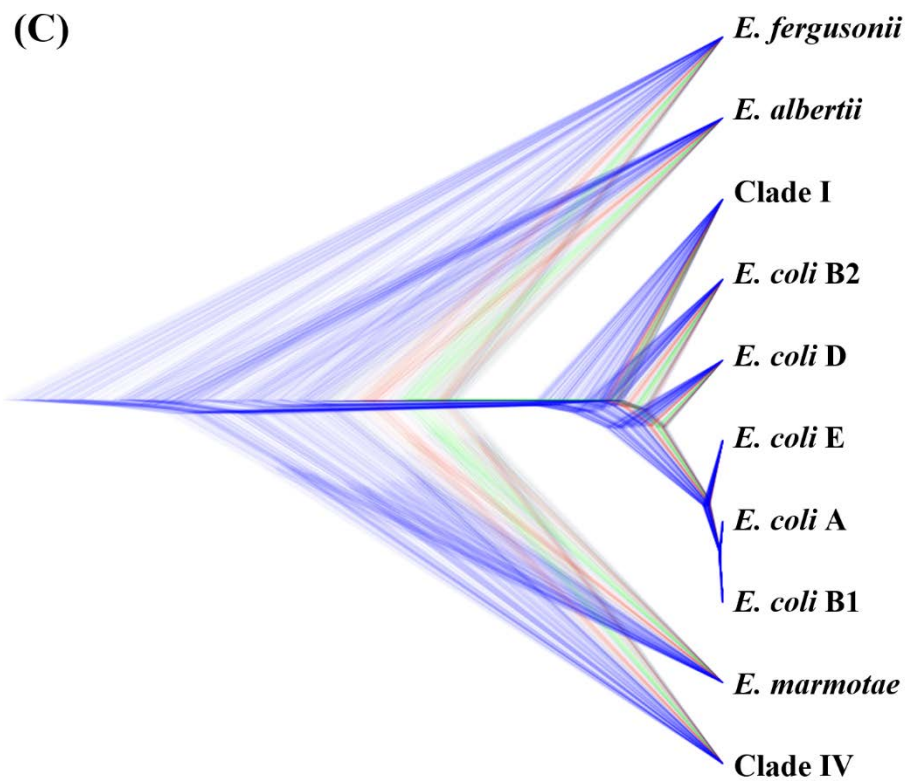

Figure 54. Continued. (C) Result from the first random subset of 60 genes randomly selected from *Enterobacteriaceae* core genes.

4.3.2. Molecular chronology of *E. coli*

Previously reports of divergence time in the evolutionary history of *E. coli* was restricted to the divergence time between *E. coli* and *S. enterica*. As genome sequences are now available for many species that diverged after the divergence of *E. coli* and *S. enterica*, it was possible to analyze the timing of more recent divergence events. Relaxed molecular clock model was applied to the 1st and 2nd codon position of the 822 core genes to estimate the divergence time for internal nodes present in the NJ tree of 6 *E. coli* strains, 5 *Escherichia* spp. other than *E. coli*, 3 *Citrobacter* spp., *3 Salmonella* spp. and 4 *Enterobacter* spp. strains. The result of time-tree analysis was depicted in **Fig. 55.**. Based on the assumption that the divergence between *E. coli* and *S. enterica* happened in between 120 MYA and 160 MYA, the earliest divergence within the genus *Escherichia* took place in between 37.9 MYA and 39.3 MYA. Subsequently, Clade IV, V and *E. fergusonii* diverged from *E. coli* and finally the species *E. coli* diverged from the clade I in between 16.6 MYA and 17.7 MYA. If clade I should be regarded as a lineage inside the species *E. coli,* then the birth of *E. coli* was 25.9 – 26.9 MYA. Assuming that this estimations were accurate, all divergences between *Escherichia* spp. took place during the Eocene, the Oligocene and the Miocene epochs of the Tertiary Period.

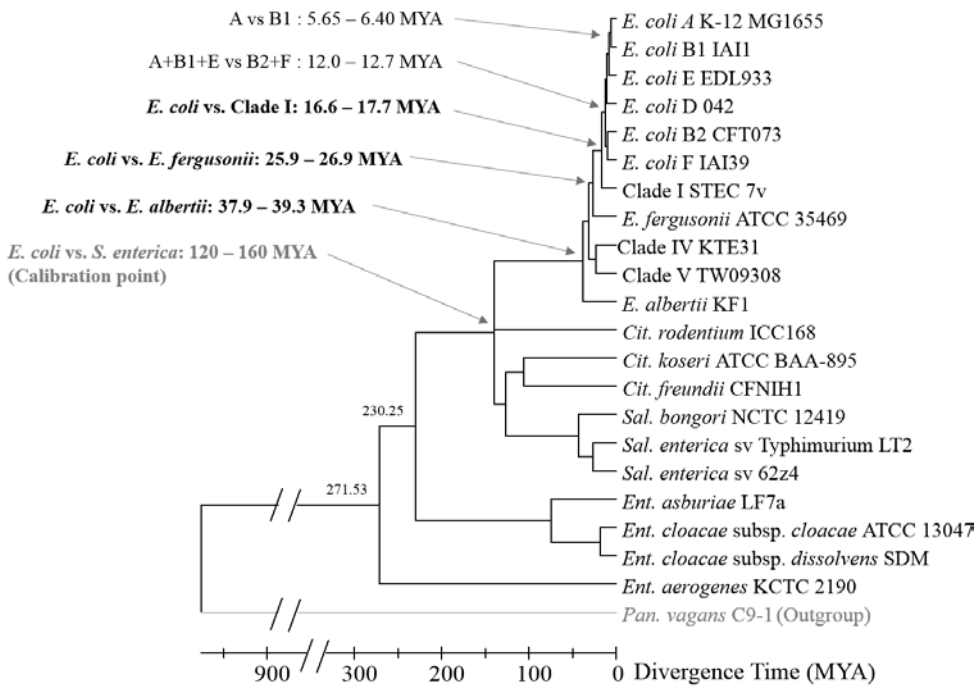

Figure 55. Timing of divergence between the species and clades of *Escherichia.*

4.3.3. Phylogenetic scenario for *Shigella* spp.

In order to evaluate phylogenetic scenario of the emergence of *Shigella* strains within *E. coli* species, the phylogenies of the large VP of *Shigella* strains were compared with the chromosomal SNP phylogeny. The plasmid phylogeny was thought to represent the history of pathogenicity evolution, because the VP has been known as a hallmark of *Shigella* pathogenicity. To collect the genes conserved in all plasmids, *E. coli/Shigella* genomes were screened for the presence of the orthologous match based on RBH relationships between the genes encoded in the large VP of *Shigella*. Genomes containing the VP were distinguished by the counts of the orthologues of VP genes. In **Fig. 56**, the genomes of the *E. coli* strains that were not designated as *Shigella* contained less than 50 RBHs of the genes of pCP301 of *S. flexneri* and pSS_046 of *S. sonnei*. A few exceptional genomes that contained >100 plasmid genes were detected. Genomes of some strains that were designated as 1 of 4 *Shigella* species contained <100 plasmid genes. The reason behind such lack of VP genes in these genomes was not investigated in this study. As visualized in the **Fig. 57**, 62 genomes that contained >100 RBHs of the VP genes (VP^+ strains) and the 83 VP core gene set that were present in 50 or more genomes were used for VP phylogeny analysis. At the same time, the chromosomal phylogeny history was reconstructed by chromosomal SNPs. While the plasmid phylogeny included only the carriers of *Shigella* VPs, the chromosomal phylogeny included the whole diversity of *E. coli*/*Shigella* strains. The resulting multigene phylogeny of 62 *Shigella* VPs and chromosomal phylogeny was compared to assess the possibilities of evolutionary scenarios (**Fig. 58**). The two phylogenies clearly denied the single-origin of *Shigella*. Even under assumption of independent entrances of the plasmid

to the 2 or 3 *E. coli* lineages, the topological differences between the two phylogenies could not be resolved. Based on my own trial to simulate the scenarios, 5 events of plasmid acquisition were required to explain the topological differences between the two phylogenies.

The plasmid phylogeny contained a monophyletic clade which was composed of the plasmids detected in the strains of *E. coli* that were not designated as *Shigella* strains. The plasmids were clustered into a monophyletic clade in the plasmid phylogeny (pink labels in **Fig. 58**), however, in the chromosomal phylogeny the strains were separated into distant branches. The clade of plasmid seemed to have been acquired by several *E. coli* lineages recently. Based on that observation, the VP of *Shigella* seem to be actively circulating in the population of *E. coli*.

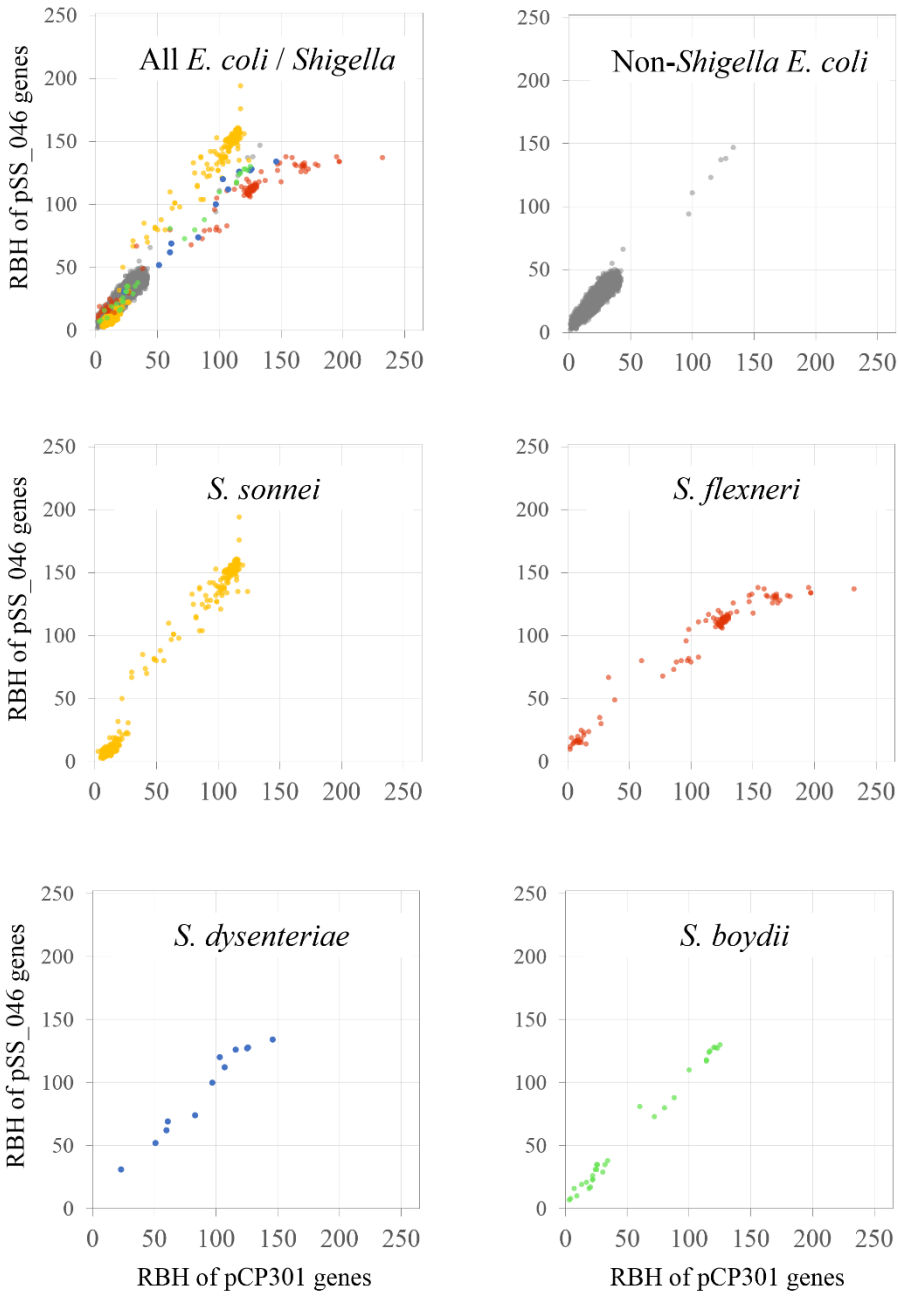

Figure 56. Occurrence of the RBHs of *Shigella* VP genes in the genomes of *E. coli* and *Shigella* strains.

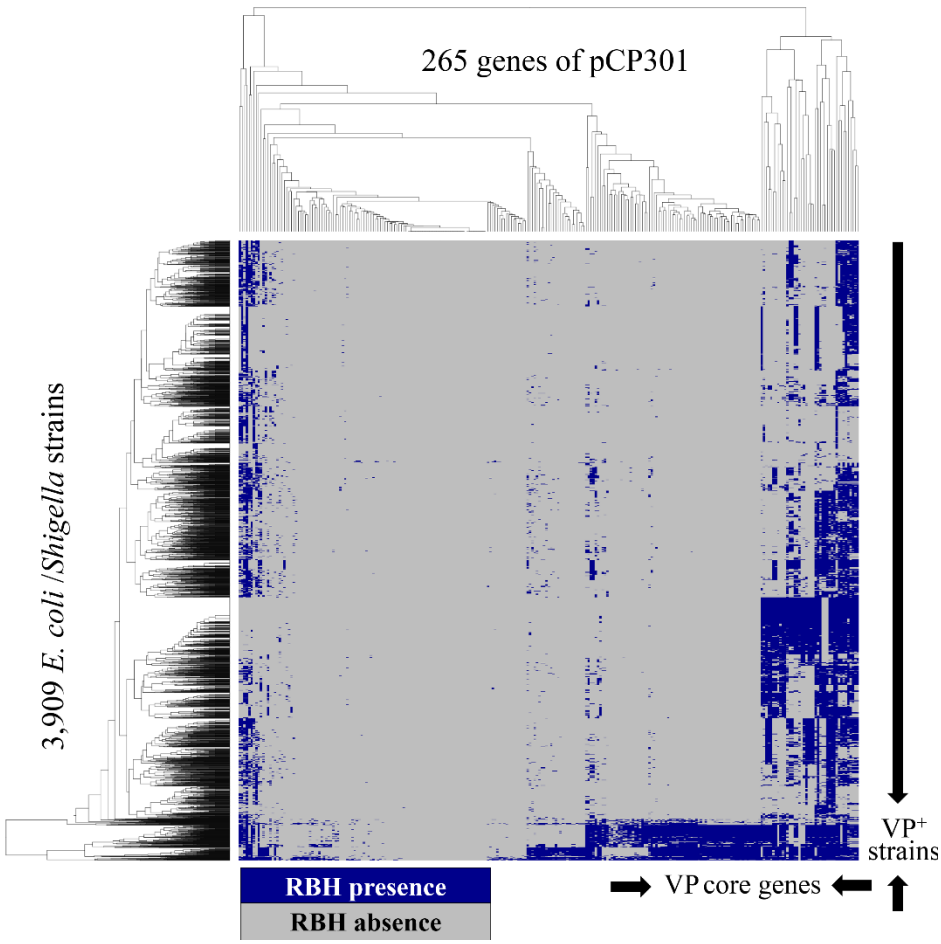

Figure 57. Selection of *Shigella* VP carrier strains and selection of the core genes of *Shigella* VPs.

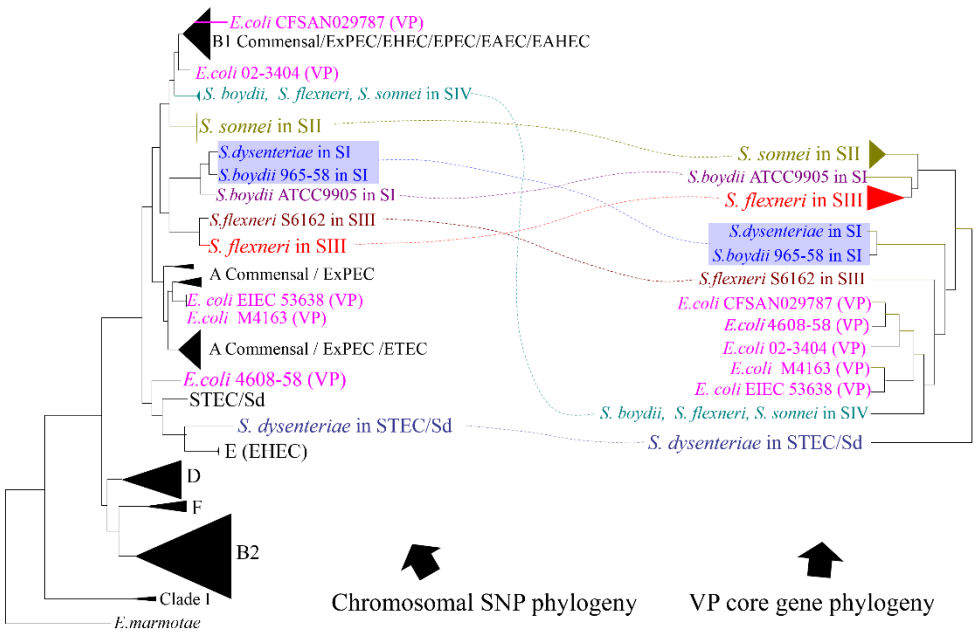

Figure 58. VP core gene phylogeny vs. chromosomal SNP phylogeny. Corresponding strains in the two trees were linked by dashed lines.

4.3.4. Genes that distinguished *E. coli* from other *Escherichia* spp.

Intermingled evolutionary history between the species in the genus *Escherichia* was demonstrated well by the results presented above. In this subchapter, the genes that distinguish *E. coli* from the other *Escherichia* spp. were analyzed. Systematic searching for species-specific core genes (i.e. the diagnostic gene sets) resulted in the discovery of 41 genes that were present in most of the *E. coli* strains and absent in the most of non-*E. coli* strains. As shown in the species tree analysis, clade I seemed to be the earliest diverged member within the species *E. coli.* Genes that distinguish *E. coli* were not found when clade I was considered as outgroup. The 41 genes selected in this study were diagnostic gene set for the strains of conventional *E. coli* plus the clade I strains. The functions of the proteins encoded by these 41 *E. coli*-defining genes were summarized in the **Table 11**. The genes in **Table 11** were particularly interesting because the genes are candidates of the driver of ecological differentiation of *E. coli* from the other species. Two operons were included in the list. The *rut* operon was known to confer the ability to catabolize the pyrimidines produced from degradation of mRNA. It has been suggested that the *rut* pathway endow advantage to the organism when the fecal to environmental dissemination cause abrupt change in the temperature and nutrient level surrounding the bacterium. The function of *phn* operon is degradation of phosphonates, a major reservoir of organic phosphorus in the environments. Other than the two operons, genes related to antibiotic resistance, transcriptional regulator and host-interaction proteins were included in the list of *E. coli* diagnostic genes.

Table 11. Functions of the gene sets that are diagnostic to *E. coli*.

Gene index	eggNOG category	Function	Note
327126444	F	Uracil permease	6 / 7 of the genes in *rut* operon
327126446	S	Predicted flavin reductase *RutF*	
327126448	C	Predicted reductase *RutE*	
327126450	S	Hydrolase or acyltransferase *RutD*	
327126452	J	Ring-opening amidohydrolase *RutC*	
327126454	Q	Predicted amidohydrolase *RutB*	
327127513	C	Molybdopterin oxidoreductase	-
327133049	NA	Metal-dependent hydrolases of the beta-lactamase superfamily I	-
327133051	S	Acetyltransferase	-
327133053	P	Guanylate kinase catalyzing phosphorylation of ribose 1,5-bisphosphate	11 / 14 of the genes in *phn* operon
327133055	P	Metal-dependent hydrolase involved in phosphonate metabolism	
327133057	P	Phosphonates transport ATP-binding protein *PhnL*	
327133060	P	Phosphonates transport ATP-binding protein *PhnK*	
327133062	P	Phosphonate metabolism protein *PhnJ*	
327133064	P	Phosphonate metabolism protein *PhnI*	
327133066	P	Phosphonate C-P lyase system *PhnH*	
327133068	P	Phosphonate C-P lyase system *PhnG*	
327133070	K	Transcriptional regulator *PhnF*	
327133075	P	Phosphonate ABC transporter phosphate-binding periplasmic component	
327133077	P	Phosphonate ABC transporter ATP-binding protein	

Table 11. Continued.

Gene index	eggNOG category	Function	Note
327129374	C	L-carnitine dehydratase	-
327129368	C	L-carnitine dehydratase	-
327129372	E	Acetolactate synthase oxalyl-CoA decarboxylase	-
327127509	K	Transcriptional regulator	-
327126561	O	Glutaredoxin 2	-
327129360	P	Multidrug transporter *MdtD*	-
327129364	T	Positive transcription regulator *EvgA*	-
327129362	V	Secretion protein *HlyD* family	-
327129366	NA	Hybrid sensory histidine kinase	-
21135500	NA	NA	-
20482658	NA	NA	-
21140184	NA	NA	-
20355844	NA	NA	-
20356165	NA	NA	-
327127511	NA	NA	-
327129378	NA	NA	-
327127515	S	Mannose-specific adhesin *FimH*	-
327128761	S	Uncharacterized protein *YegR*	-
327129376	S	*YfdX* protein	-
327129370	S	Auxin efflux carrier	-
327126804	S	Prophage protein	-

Given that the two operons, *rut* and *phn* were not present in the other *Escherichia* spp., the evolutionary origin of the two operons in *E. coli* was of great interest. The homologs of all gene in the operons were searched for in the entire prokaryotic genomes. The intact set of *rut* genes were found in *Enterobacter, Klebsiella, Leclercia* and *Staphylococcus.* The intact gene set of *phn* operon was found in *Enterobacter, Citrobacter, Leclercia, Salmonella, Pseudomonas* and *Staphylococcus.* The two operons had significant overlap in the taxonomic distribution of their homologs. Phylogenetic analysis of the genes of the operons resulted in the trees shown in **Fig. 59** for *rut* operon and **Fig. 60** for *phn* operon. In **Fig. 59** *rut* operons of *E. coli* formed a tight cluster separated from the other species. One *K. pneumoniae* strain was found within the *E. coli* clade. In this case, the original o donor of this operon that gave this operon to the ancestor of *E. coli* cannot be identified based on this phylogeny. Likely, in **Fig. 60** the *E. coli phn* sequences formed a distant clade that was not affiliated in the clade of other species. Therefore, the donor of *phn* operon cannot be identified by this phylogenetic analysis. However, interestingly a *K. pneumoniae* strain, a *S. hominis* strain and an unclassified *Enterobacteriaceae* strain were affiliated to the branches inside the clade of *E. coli phn* genes. These cases are more likely resulted from the transfer of operon from *E. coli* to those species.

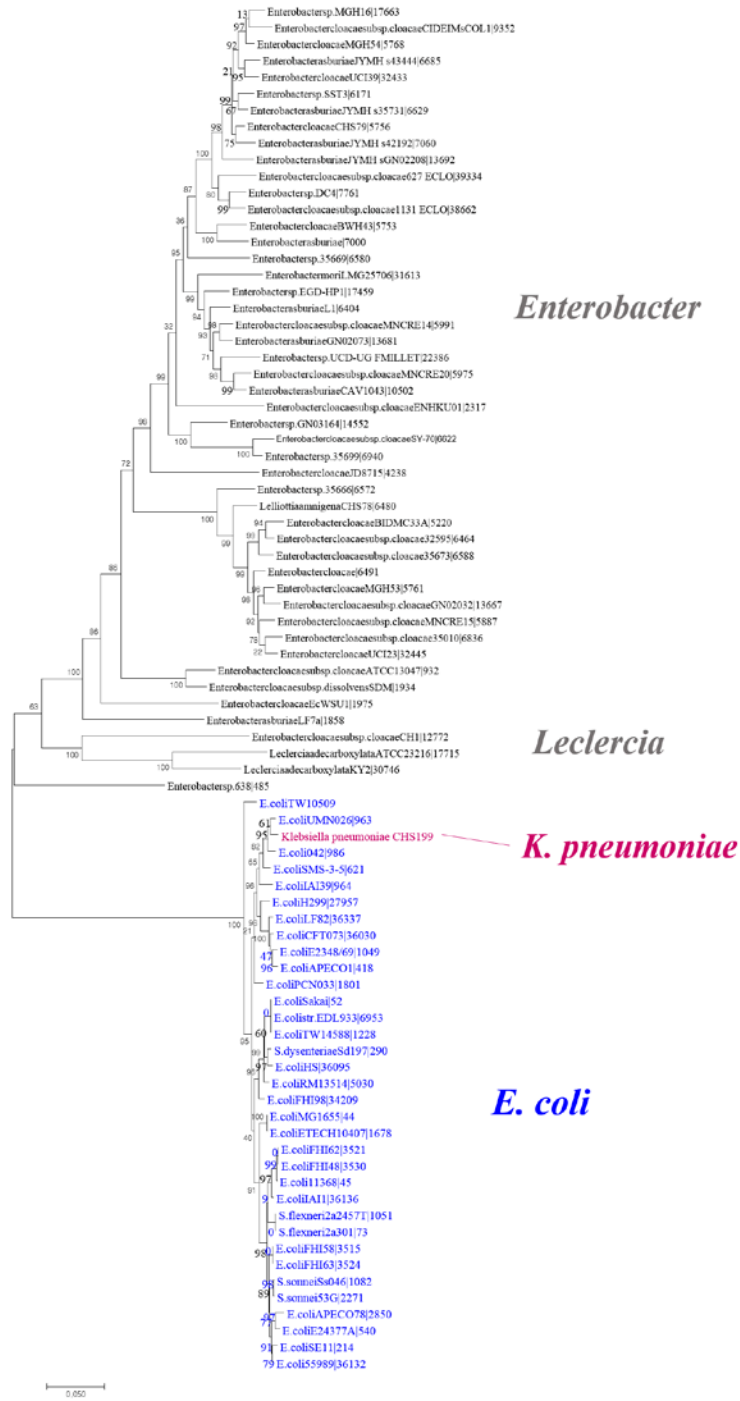

Figure 59. Multi-gene phylogeny of the *rut* operon of *E. coli* and its homologs.

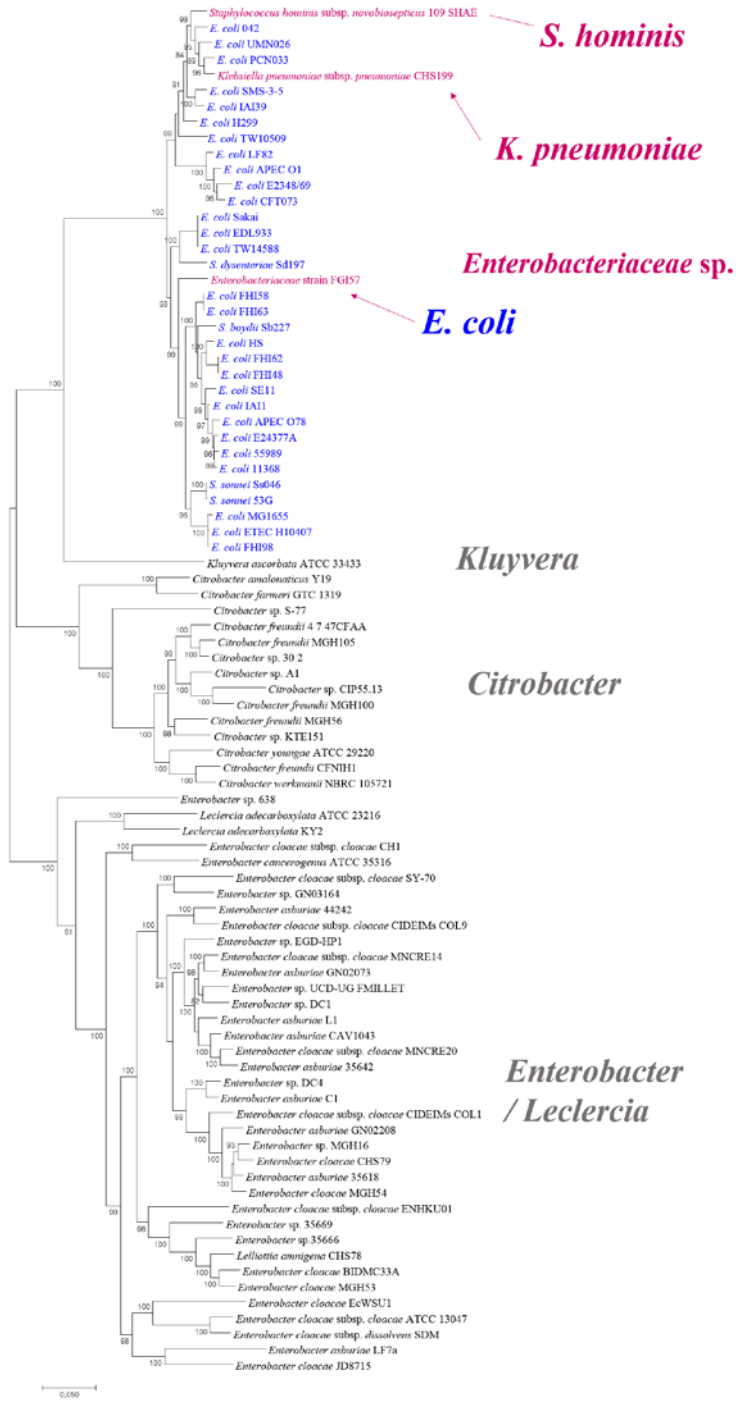

Figure 60. Multi-gene phylogeny of the *phn* operon of *E. coli* and its homologs.

4.4. Discussion

In the chapter 4 genome evolution of *E. coli* was analyzed in the context of evolution at larger scale, focusing on the relationship of *E. coli* and the neighboring species. Analyses were based on the comparison of the genomes representative of *E. coli* and the closely related species. Phylogenomic network of the species that are members of 21 genera in the family *Enterobacteriaceae* indicated that *Enterobacteriaceae* have diverged by radiation, with extensive genetic exchanges in the early stage of diversification. A similar pattern was nested in the phylogenomic network of *Escherichia* spp. strains. Support for the presence of active horizontal genetic exchanges came from the presence of abundant horizontal branches in the phylogenetic networks. Indication of divergence by radiation came from the branching pattern which contained long branches leading to the extant species that are inter-connected by short branches in the basal area (i.e. star phylogeny). The indicated extensive genetic exchange at the basal area of phylogeny was coherent with what could be expected under the hypothesis of temporally fragmented speciation (Retchless and Lawrence 2007). The hypothesis of temporally fragmented speciation of bacterial genomes claimed that different genomic regions have diverged independently at different timing (Retchless and Lawrence 2007, 2010). Based on the hypothesis, in the evolutionary history of *Enterobacteriaceae*, some genomic regions have diverged in the earlier stage of speciation and recombination between the populations (that are becoming distinct species) have stopped in those genomic regions. In the other genomic regions genetic exchange by recombination was maintained for a period of time until speciation took place later in those regions.

Under that hypothesis, extensive horizontal lines between the species that are currently divergent should be detected in the early phase of *Enterobacteriaceae* diversification. The pattern of explosive radiations is considered as a general pattern of the evolutionary history of macroorganisms (Morlon, et al. 2010). For microorganisms temporal patterns of diversification has been discussed just by less than a handful of few studies, and the conclusions from different studies did not agree on the occurrence of radiation. One previous study that covered diverse lineages of *Acidobacteria*, *Crenarchaeota*, *Haloarchaea* and Methanogens concluded that the diversification of these groups of organisms occurred in constant rates and not by adaptive radiation (Martin, et al. 2004). A study that tested the evolutionary history of *Borrelia burgdorferi* sensu lato concluded that an explosive radiation of lineages occurred near the origin of this species group (Morlon, et al. 2012). A study of temporal diversification of the genus *Aeromonas* concluded that the temporal rates of diversification in *Aeromonas* was constant over time and positively correlated with the number of animal genera (Lorén, et al. 2014). The phylogenomic analysis of *Enterobacteriaceae* in this chapter did not include the direct estimation of past diversification rates, however, based on the patterns of branching it seems that adaptive radiation occurred at the origin of *Enterobacteriaceae* and at the origin of *Escherichia.* To corroborate this hypothesis, explicit estimation of the branching rates over time should be performed in the further study. Like in case of *Aeromonas*, the diversity of animal hosts is expected to be correlated with the temporal diversification of *Escherichia* and *Enterobacteriaceae.*

The timing of the origin of *E. coli* has not been estimated with as much genomic data as what is available now. The previous estimation of divergence time

was made for the common ancestor of *E. coli* and *S. enterica* serovar Typhimurium. The two species are not the closest relative to each other and therefore the estimated age of their common ancestor did not convey the age of *E. coli.* By including the closest relatives of *E. coli* in the time-tree analysis it was able to narrow down the origin of *E. coli* to 16.6 – 17.7 MYA. The previously reported 120 MYA (the divergence time between *E. coli* and *S. enterica* serovar Typhimurium) and the newly reported 17 MYA are drastically different if animal fauna on the earth is considered, as the former corresponds to the early Cretaceous and the latter corresponds to the early Miocene. Based on the studies of mammalian diversification history, major lineages including the *Rodentia*, *Primates*, *Carnivora* and *Chiroptera* all originated almost simultaneously around 85 – 92 MYA (Bininda-Emonds, et al. 2007; Stadler 2011). As the estimated divergence time of *E. coli* in this study put the origin of *E. coli* after the origins of diverse mammalian hosts of *E. coli*, it is not likely that an ancestor of *E. coli* was present in the mammalian ancestor and the host range of *E. coli* has been expanded as a result of the diversification of mammals. In other word, such a simple explanation for the origin of the wide host range of *E. coli* should be excluded. It would be interesting to inquiry if the species *E. coli* had wide host range from the beginning, or had a narrow host range at first and expanded later. Exact host range of *E. coli* has not been fully explored, however, a couple of large scale surveys of *E. coli* genetic structure in diverse environments were carried out. In a survey of 2,300 vertebrate animals in Australia, *E. coli* was showed the highest prevalence in mammals, followed by birds and was negligible in fish, frogs and reptiles (Gordon and Cowling 2003). A study of 162 Amerindians and 198 animals in French Guiana showed that the prevalence of *E. coli* was 100% in humans, 64% in domestic animals and 45% in wild animals (Lescat, et al. 2013). A study that screened the frequency

of *Escherichia* clades in >3,500 isolates from diverse fecal samples suggested that the *Escherichia* clades other than *E. coli* had low frequency in human samples (2-3%) and in other mammals (3-8%) and had higher frequency in birds (8-28%) (Clermont, et al. 2011). Differential composition of phylogenetic groups of *E. coli* between human isolates, non-human mammalian isolates and bird isolates was also reported (Escobar-Páramo, et al. 2006). Results of these previous studies roughly suggested that *E. coli* and the other clades of *Escherichia* had differential host preference. Based on this study the divergence within the genus *Escherichia* occurred during 38 MYA – 17 MYA, and most lineages of modern mammals and birds were generally present in that epoch (Claramunt and Cracraft 2015). Differential host preferences may have been the driver of evolutionary divergence between the clades in *Escherichia.* For example, the ability to attach to the epithelium of human gut is mediated by the secretion systems and pili (de Muinck, et al. 2013). In the previous chapter, the secretion proteins and fimbrial proteins were frequently detected as a target of positive selection, the sign of adaptive evolution. Combining all the indications with the timing of divergence estimated in this chapter, there is a probability that the adaptive evolution at the locus that mediate host preference have played a central role in the divergence of the clades in *Escherichia.*

Evolutionary relationship between *E. coli* and the four species of *Shigella* has been discussed in numerous previous studies based on whole genome analyses (Yang, et al. 2007; Peng, et al. 2009; Sims and Kim 2011; Zhang and Lin 2012; Gordienko, et al. 2013; Sahl, et al. 2015) and MLST data (Pupo, et al. 2000). One unambiguous conclusion from the previous studies was that the *Shigella* species are clones of *E. coli*. However, consensus has not been made regarding the detailed phylogenetic

scenario for the emergence of *Shigella* species. In the previous studies of gene composition and genomic feature frequencies, *Shigella* genus formed a single clade in terms of phenetic phylogeny (Ogura, et al. 2009; Sims and Kim 2011). In contrary, studies of the sequences at the conserved genomic loci have argued that *Shigella* genus appeared in several independent occasions and the exact numbers of origins were variable between the studies. The earliest study based on a few number of housekeeping genes study argued that *Shigella* arose in at least 7 independent branches (Pupo, et al. 2000). A following study based on the housekeeping genes and 4 virulence factor genes insisted on a single phylogenetic origin of all *Shigella* strains (Escobar-Páramo, et al. 2003). Later studies of shared core-genomic loci claimed at least 3 occasions (Zhang and Lin 2012) and 5 monophyloetic clades (Sahl, et al. 2015). The similarity of genomic compositions and phenotypic characters of *Shigella* strains and the polyphyletic nature of the genus led to the paradigm of parallel convergent evolution of this pathogens (Maurelli 2007). In this chapter, the phylogenetic tree of chromosomal SNPs and the phylogenetic tree of the core genes of *Shigella* VPs were compared to illustrate the phylogenetic scenario for acquisitions of the VPs. The chromosomal SNP phylogeny was assumed to represent the vertical evolutionary history of the majority of genomic backbone, while the plasmid phylogeny was assumed to reflect the evolutionary history of VP that were frequently transferred horizontally within the *E. coli* population. An apparent pattern was that among *E. coli* phylogenetic groups the clade consisted of group B2, D and F were free from the acquisition of *Shigella* VPs. All acquisition events occurred in the other clade (group A, B1, C and E) and happened all across the sub-branches. The two phylogenies were highly incongruent and the reconstruction the number of acquisition events were a puzzling guesswork. The exact number of independent

gains of the plasmid could only be inferred in terms of the minimum required number of events. Without supposing 5 or more independent acquisition events, it was unable to reconcile the two phylogenies. In the course of extracting the plasmid genes from *E. coli* genomes, a novel clade of VPs was discovered in the strains that were not described as *Shigella*. The strains containing the plasmids that belonged to this novel clade were found in 4 different chromosomal lineages and were distributed across group A, B1 and E. This observation demonstrated that the lateral spread of the plasmid could be happening in high pace. Clinical symptoms might not be manifested by the strains. In the recent theories of patho-adaptive evolution of *Shigella* genomes, significant genome reduction by loss of genes is a convergent process in all *Shigella* lineages, that was essential for successful switching into highly specific pathogenic lifestyle (Hershberg, et al. 2007). With that in mind, acquisition of VPs should be regarded as a driver that obligate the new lifestyle, not as a sign of finalization of the *Shigella* lifestyle.

Systematics studies of *E. coli* corroborated the idea that the species within the genus *Escherichia* had an intermingled evolutionary history. As exemplified by the inclusion of the genus *Shigella* inside the *E. coli*, the species *E. coli* harbors remarkable ecological and phenotypic diversity. Given such a complex history and diverse lifestyles, it was interesting to search for the genomic features that unify the *E. coli*/*Shigella* species group. In the previous chapter, 41 genes that were characteristic to *E. coli* and not present in the other *Escherichia* spp. were discovered. The functions of these species-specific unique core genes might offer insights about the nature of ecological differentiation that distinguished *E. coli* from its relatives from the early stage of its evolution. A similar approach in the *Vibrionaceae* resulted

in the suggestion that the unique functions of *V. cholerae*, aerotaxis and vibriobactin (iron chelation), provided the driver of adaptive divergence of *V. cholerae* (Kahlke, et al. 2012). Among the 41 unique core genes of *E. coli*, two operons comprised 17 genes. Six of seven genes in the *rut* operon for the novel pyrimidine catabolism pathway discovered in *E. coli* were found to be the species-specific core genes. The pyrimidine catabolism pathway was shown to make the cell capable of assimilating nitrogen by degrading the pyrimidines generated from turned-over RNA. Furthermore the operon encodes for a putative uracil transporter, thus have a potential to uptake environmental pyrimidine molecules (Parales and Ingraham 2010). This pyrimidine catabolism pathway was found in other soil-residing species of the phylum *Proteobacteria*. The pathway is thought to confer a metabolic benefit in the situation of down-ward shift of temperature and nutritional level in the environment, such as the shedding of fecal microorganisms into the environments (Kim, et al. 2010). Another operon that was specific to *E. coli* encoded for the C-P lyase pathway, the predominant route for phosphonate utilization in the microbes. The *phn* operon consists of 14 genes and 13 of them were found to be *E. coli*-specific in this analysis. Phosphonates are chemicals that contain C-P bond and are known to be a major reservoir of dissolved organic phosphorus. C-P lyase pathway of *E. coli* was shown to be capable of catalyze diverse forms of phosphonates, through the dephosphonation reactions that are generally difficult to occur (Villarreal-Chiu, et al. 2012). The ecological role of this pathway has been studied mostly in marine microbes where it is found frequently. It is quite interesting to observe that the two operons that distinguished *E. coli* from the other *Escherichia* spp. provided the ability to acquire nitrogen and phosphate from unusually utilized source of them. Analysis of phylogenetic relationships of *E. coli* rut operon genes and its homologs

in other bacteria revealed that the operon was mostly confined to the *Enterobacteriaceae* species. The *rut* operon of *E. coli* strains formed a monophyletic clade and were closely related with that of *Enterobacter* strains. The homologs of *E. coli phn* operon were found in a taxonomic distribution that is similar with that of *rut* operon. The multi-gene phylogeny of the *phn* operons, like that of *rut* operon, indicated that the gene cluster of *E. coli* is close with that of *Enterobacter* strains. It is unclear whether the similarities of the taxonomic compositions of the homologs of the two operons were made by chance or by certain ecological reason. While the two operons discussed above mediate the basic nutritional advantages of *E. coli*, other *E. coli*-specific genes encoded for diverse functions. Among them, notable genes with known functions are a beta-lactamase gene, a multidrug transporter *mdtD*, an *L*-carnitine dehydratase gene, a glutaredoxin 2 gene, a transcription regulator *evgA*, oxalyl-CoA decarboxylase gene, *fimH* adhesin and a type I secretion protein *hlyD*. Of them, detection of *fimH* and *hlyD* as species-specific unique core genes was interesting because the genes contribute to the pathogenicity of the pathogenic clones of *E. coli* (Lenders, et al. 2015; Sauer, et al. 2016). Unlike the basic nutritional roles of nitrogen and phosphorus assimilation pathways discussed above, the roles of adhesion and secretion are related with the interaction with the hosts. Combined all, the detected *E. coli*-specific core genes contained a variety of roles that are potentially beneficial to the lifestyle in which the organisms experience both the intra-host environments and the outside environments.

CHAPTER 5

Conclusions

In this thesis, phylogenetic analysis and pan-genome analysis were combined to decipher the patterns in *E. coli* genomic diversity. The time-scale regarded in this thesis is not as short as that of epidemiological studies where the evolution within the clones of bacteria that happens within a hundred years is covered, and not as long as that of 'tree of life' studies that deal the divergence between multiple families of species. The results from the examination of evolutionary processes that happened inside the species of *E. coli* suggested that the formation of phylogenetic groups were accompanied by the shift from sexual to clonal mode of evolution. Evidences from multiple analyses also suggested that the rate of evolutionary diversification of core-genome sequences are dissociated from the rate of diversification of gene repertoire of the strains. Combined together, our observation of relatively higher rates of pan-genome diversity per unit phylogenetic diversity in the phylogenetic groups, compared to that of the species *E. coli*, might be explained by the repression of core-genome sequence diversification in the clonal genomes. In the following paragraphs each line of the above arguments were discussed in detail.

Presence of intra-specific phylogenetic structure of *E. coli* has been demonstrated in a number of previous studies (Touchon, et al. 2009; Walk, et al. 2009; Lukjancenko, et al. 2010; Tenaillon, et al. 2010). Discussions about the nature of the presence of phylogenetic groups within *E. coli* are connected to the nature of evolutionary processes that shape the bacterial species. A hypothesis of recombinational dormancy of *E. coli* and a hypothesis of on-going speciation between the phylogenetic groups of *E. coli* provided the theoretical idea tested in this thesis. In the hypothesis of recombinational dormancy of *E. coli* (Leopold, et al. 2011) it was argued that exchanges of DNA between phylogenetic groups were

largely inactivated in the extant *E. coli* population. Inactive genetic exchange between the groups was regarded as a sign that the entity as a species is no longer maintained for *E. coli*. On-going speciation between *E. coli* phylogenetic groups was hypothesized again based on the estimation of the frequencies of inter- and intra-phylogroup gene flows by recombination (Didelot, et al. 2012). In their study, group A+B1, group E and group B2 each displayed relatively low inter-group recombination activity compared to the intra-group recombination activity. The biased pattern in genetic exchange was regarded as a sign of on-going process that resembled sexual isolation of bisexual species. According to the core-genome phylogeny reconstructed in this study the strains of *E. coli* were able to be classified into known monophyletic clades plus a few miscellaneous clades. Multiple results made in this study suggested that the presence of stable and distinct phylogenetic clades within *E. coli* is related with the decline of recombination rates through evolutionary history of *E. coli*. First, from branch-specific estimation of relative rates of recombination-driven base changes and mutation-driven base changes, the impact of recombination was lower for recent branches than for deep branches. Second, prevalence of recombined genes determined by the analysis of each individual phylogenetic groups were lower, compared to what was observed for the entire *E. coli* strains. Third, the decay of linkage between SNP sites was weaker within the strains of the same phylogenetic group. Finally, branching patterns of intra-species phylogeny of *E. coli* displayed the characteristic patterns of sexual bacterial species in the basal region, while more characteristic patterns of clonal bacterial species were observed in the tip area of the phylogeny. Estimation of inter-group recombination rates was not in conflict with the previously published on-going speciation hypothesis (Didelot, et al. 2012), as the group B2, D and F exhibited the strongest

internal gene flow and the group B1+C and group A each received the highest gene flow from the group B1+C.

Quantification of the genomic diversification of *E. coli* was dissected into the analysis of DNA sequence diversification within each gene and the diversification of gene repertoire contained in the genomes. An emerging pattern from several analyses was that the pace of two processes could be uncoupled from each other. Characteristics of sequence evolution in the core genes and rare genes were contrasted in two aspects. Core genes showed more narrowly confined distribution of the level of polymorphism, while phyletically rare genes showed wide range of polymorphism, from extremely conserved to highly diverse. Selective pressures acting on the core genes were much more biased to negative purifying selection than those acting on phyletically rare genes. The concept that the rate of sequence diversification in the core genes was dissociated from the rate of gene content diversification was indicated from several findings. Firstly, it was observed that while sequence divergence has not accumulated without gene content divergence, gene content divergence did accumulate without sequence divergence to some degree. Second, investigation of gene orders in many strains revealed that the synteny of genes were conserved throughout entire core-genome even in the presence of severe in-out flux of genes. Finally, comparison of the bacterial species that take diverse positions in the spectrum from clonality to sexuality, the shape of pan-genome gene frequency distribution was universally preserved. The first and second findings were direct evidence of dissociation between core-genome evolution and pan-genome evolution, while the last finding provided a circumstantial evidence for it. Under neutral evolution, the diversification of sequence and gene content

would be achieved at constant rates and not uncoupled from each other. In the presence of natural selection, the rate of sequence and/or gene content diversification could be deviated from their neutral rates. Natural selection on sequences could act to slower the pace of sequence diversification by several mechanisms. Negative selection by definition would repress the diversification of sequences and positive selection could purge the diversity under clonal setting. The impact of natural selection on the variation of gene contents is largely unknown at this moment. Various statistical models and methods of inference are available to identify the natural selection affecting DNA sequence evolution, but the models for natural selection's effect on the gene content turnover has not been established yet, albeit a few models have been suggested and discussed (Baumdicker, et al. 2012; Lobkovsky, et al. 2013; Lobkovsky, et al. 2014). Using the available method, at least on DNA sequence level, the presence of both positive and negative selection was confirmed in *E. coli* genomes.

In addition to reproducing the observation of open-pan-genome of *E. coli*, analysis of pan-genome in this study indicated that the relative rate of pan-genome expansion and core-genome sequence diversification was higher for each phylogenetic group, when compared to that of *E. coli* species. Conclusions from the above two paragraphs were (i) that *E. coli* evolution has shifted toward clonal end and (ii) the pace of gene content diversification is dissociated from the pace of core-genome sequence diversification. The two conclusions can be combined to explain the given result that the rate of gene content diversification per unit phylogenetic diversity was higher for *E. coli* sub-population groups than for *E. coli* species. Sequence diversification could be slowed down as genomes become more clonal, by

the process of genome-wide selective sweep by periodic selection which results in the genome-wide purging of sequence diversity (Cohan 2001; Takeuchi, et al. 2015; Bendall, et al. 2016). The process of selective sweep has been discussed for long time based on theoretical simulations until recently, however, a recent study using metagenome sequences obtained from a freshwater lake provided the evidence that such a process actually takes place in the natural bacterial populations (Bendall, et al. 2016). The efficiency of selective sweep is expected to be stronger in the population when the genome-wide linkage level is higher. If periodic selection was occurring in the population of *E. coli*, the impact of erasing sequence polymorphisms would be higher within the phylogenetic groups, since the genome-wide linkage level was shown to be higher within the phylogenetic groups. A condition required for periodic selection to happen is the occurrence of advantageous mutation. Analysis of dN/dS ratio over pan-genome of *E. coli* indicated that positive selection was occurring rarely in the core genes, and with higher frequency in non-core genes. When the analysis of dN/dS ratio was confined to the sequences within the same phylogenetic groups, the prevalence of positive selection was observed more frequently. Therefore, positive selection of novel mutations does happen in *E. coli* genomes, particularly more so within each phylogenetic group. In conclusion, elevated linkage level and frequencies of positive selection in the phylogenetic groups could be regarded as that the conditions required for periodic selection to take place were provided in the phylogenetic groups of *E. coli.* Unfortunately direct evidence of periodic selection is not presented in this thesis, and further analysis should be required. Assuming that, the elevated rate of pan-genome expansion per unit increase of phylogenetic diversity could be resulted from deceleration of sequence divergence within clonal subpopultions.

Observation of decline of recombination rates and shifting from recombining to clonal genetics in *E. coli* was a major part of the thesis. Elucidation of the reasons, the ecological, demographic and evolutionary backgrounds of such historical trend certainly is a more important step than the discovery of such trend. Extensive analysis should be performed using ecological metadata, palaeoecological data, demographic analysis, more sophisticated analysis of natural selection, and more systematically sampled genome sequences. Moreover, to address the evidences of adaptive evolutionary processes in the natural population, analysis of genome sequence data might not be sufficient. The population level gene frequency and/or allele frequency data would be required to provide evidences of adaptive evolution, as seen in human genetics studies. The scheme of analyzing randomly collected genomes cannot provide the dynamics of gene/allele frequencies at population level. In personal opinion a promising solution to this problem can be found in the metagenomics data. Accumulation of microbiome data already has opened the chance to analyze local genetic structure and population dynamics of diverse bacterial species.

REFERENCES

Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. *Nature Review Microbiology* 6:431-440.

Amaya E, Reyes D, Paniagua M, Calderón S, Rashid MU, Colque P, Kühn I, Möllby R, Weintraub A, Nord CE. 2012. Antibiotic resistance patterns of *Escherichia coli* isolates from different aquatic environmental sources in Leon, Nicaragua. *Clinical Microbiology and Infection* 18:E347-E354.

Andam CP, Gogarten JP. 2011. Biased gene transfer in microbial evolution. *Nature Review Microbiology* 9:543-555.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263-265.

Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology* 4:1-14.

Baumdicker F, Hess WR, Pfaffelhuber P. 2012. The infinitely many genes model for the distributed genome of bacteria. *Genome Biology and Evolution* 4:443-456.

Bendall ML, Stevens SLR, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, et al. 2016. Genome-wide

selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* 2016:1-13.

Bennett GM, Abbà S, Kube M, Marzachì C. 2016. Complete genome sequences of the obligate symbionts "*Candidatus* Sulcia muelleri" and "*Ca.* Nasuia deltocephalinicola" from the pestiferous leafhopper macrosteles quadripunctulatus (*Hemiptera: Cicadellidae*). *Genome Announcements* 4: e01604-15.

Bennett GM, Moran NA. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution* 5:1675-1688.

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417:141-147.

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* 446:507-512.

Bobay L-M, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. *Proceedings of the National Academy of Sciences* 112:8893-8900.

Bohlin J, Brynildsrud OB, Sekse C, Snipen L. 2014. An evolutionary analysis of genome expansion and pathogenicity in *Escherichia coli. BMC Genomics* 15:1-13.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10:e1003537.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26:1372-1373.

Brüssow H, Canchaya C, Hardt W-D. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* 68:560-602.

Brenner DJ, Davis BR, Steigerwalt AG, Riddle CF, McWhorter AC, Allen SD, Farmer JJ, Saitoh Y, Fanning GR. 1982. Atypical biogroups of *Escherichia coli* found in clinical specimens and description of *Escherichia hermannii* sp. nov. *Journal of Clinical Microbiology* 15:703-713.

Brenner DJ, McWhorter AC, Knutson JKL, Steigerwalt AG. 1982. *Escherichia vulneris*: a new species of *Enterobacteriaceae* associated with human wounds. *Journal of Clinical Microbiology* 15:1133-1140.

Bruen T. C., Philippe H., Bryant D. 2006. A simple robust statistical test for detecting the presence of recombination. *Genetics* 172:2665-2681**.**

Bryant D, Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21:255-265.

Byappanahalli M, Fowler M, Shively D, Whitman R. 2003. Ubiquity and persistence of *Escherichia coli* in a midwestern coastal stream. *Applied and Environmental Microbiology* 69:4549-4555.

Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. 2009. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proceedings of the National Academy of Sciences* 106:12412–12417.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* 2:e383.

Chun J, Rainey FA. 2014. Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. *International Journal of Systematic and Evolutionary Microbiology* 64:316-324.

Claramunt S, Cracraft J. 2015. A new time tree reveals Earth history's imprint on the evolution of modern birds. *Science Advances* 1:e1501005.

Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports* 5:58-65.

Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. 2011. Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. *Environmental Microbiology* 13:2468-2477.

Cohan FM. 2001. Bacterial species and speciation. *Systematic Biology* 50:513-524.

Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Current Biology* 17:R373-R386.

Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine. *Bioinformatics* 26:2914-2915.

Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* 31:587-589.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* 79:7696-7701.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology* 13:e1002112.

Darmon E, Leach DRF. 2014. Bacterial genome instability. *Microbiology and Molecular Biology Reviews* 78:1-39.

de Muinck EJ, Lagesen K, Afset JE, Didelot X, Rønningen KS, Rudi K, Stenseth NC, Trosvik P. 2013. Comparisons of infant *Escherichia coli* isolates link genomic profiles with adaptation to the ecological niche. *BMC Genomics* 14:1-21.

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.

Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Computational Biology* 11:e1004041.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome Research* 19:744-756.

Dubey AK, Baker CS, Suzuki K, Jones AD, Pandit P, Romeo T, Babitzke P. 2003. *CsrA* regulates translation of the *Escherichia coli* carbon starvation gene, *cstA*, by blocking ribosome access to the *cstA* transcript. *Journal of Bacteriology* 185:4450-4460.

Dykhuizen DE, Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *Journal of Bacteriology* 173:7257-7268.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461.

Escherich T, Bettelheim KS. 1988. The intestinal bacteria of the neonate and breast-fed infant. *Reviews of Infectious Diseases* 10:1220-1225.

Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E. 2003. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *Journal of Molecular Evolution* 57:140-148.

Escobar-Páramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, Skurnik D, Denamur E. 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environmental Microbiology* 8:1975-1984.

Farmer JJ, Fanning GR, Davis BR, O'Hara CM, Riddle C, Hickman-Brenner FW, Asbury MA, Lowery VA, Brenner DJ. 1985. *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of *Enterobacteriaceae* isolated from clinical specimens. *Journal of Clinical Microbiology* 21:77-81.

Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J, Dougherty B, Merrick J, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496-512.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323:741-746.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. *Science* 315:476-480.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397-404.

Gordienko EN, Kazanov MD, Gelfand MS. 2013. Evolution of pan-genomes of *Escherichia coli, Shigella spp., and Salmonella enterica*. *Journal of Bacteriology* 195:2786-2792.

Gordon DM, Cowling A. 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* 149:3575-3586.

Goto DK, Yan T. 2011. Genotypic diversity of *Escherichia coli* in the water and soil of tropical watersheds in Hawaii. *Applied and Environmental Microbiology* 77:3988-3997.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307-321.

Guttman D, Dykhuizen D. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266:1380-1383.

Han K, Li Z-f, Peng R, Zhu L-p, Zhou T, Wang L-g, Li S-g, Zhang X-b, Hu W, Wu Z-h, et al. 2013. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. *Scientific Reports* 3:2101.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han C-G, Ohtsubo E, Nakayama K, Murata T, et al. 2001. Complete genome sequence of enterohemorrhagic *Eschelichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Research* 8:11-22.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27:570-580.

Hershberg R, Tang H, Petrov DA. 2007. Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biology* 8:R164.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research* 44:D286-D293.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. *Nature Microbiology* 1:16048.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23:254-267.

Huys G, Cnockaert M, Janda JM, Swings J. 2003. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *International Journal of Systematic and Evolutionary Microbiology* 53:807-810.

Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ. 2006. Presence and growth of naturalized *Escherichia coli* in temperate soils from lake Superior watersheds. *Applied and Environmental Microbiology* 72:612-621.

Jang J, Di DYW, Lee A, Unno T, Sadowsky MJ, Hur H-G. 2014. Seasonal and genotypic changes in *Escherichia coli* phylogenetic groups in the Yeongsan river basin of South Korea. *PLoS ONE* 9:e100585.

Jombart T, Devillard S, Balloux F.. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11:1-15.

Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD. 2011. Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. *Biology Direct* 6:1-16.

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* 13:1-13.

Kahlke T, Goesmann A, Hjerde E, Willassen NP, Haugen P. 2012. Unique core genomes of the bacterial family *Vibrionaceae*: insights into niche adaptation and speciation. *BMC Genomics* 13:1-12.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44:D457-D462.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30:772-780.

Kim K-S, Pelton JG, Inwood WB, Andersen U, Kustu S, Wemmer DE. 2010. The *Rut* pathway for pyrimidine degradation: novel chemistry and toxicity problems. *Journal of Bacteriology* 192:4089-4102.

Kim M, Oh H-S, Park S-C, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 64:346-351.

Koonin EV. 2003. Horizontal gene transfer: the path to maturity. *Molecular Microbiology* 50:725-727.

Koonin EV, Wolf YI. 2012. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Frontiers in Cellular and Infection Microbiology* 2:119.

Korea C-G, Badouraly R, Prevost M-C, Ghigo J-M, Beloin C. 2010. *Escherichia coli* K-12 possesses multiple cryptic but functional chaperone–usher fimbriae with distinct surface specificities. *Environmental Microbiology* 12:1957-1977.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genetics* 4:e1000304.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5:1-9.

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, et al. 2015. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* 15:141-161.

Lee I, Ouk Kim Y, Park S-C, Chun J. 2016. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology* 66:1100-1103.

Lenders MHH, Weidtkamp-Peters S, Kleinschrodt D, Jaeger K-E, Smits SHJ, Schmitt L. 2015. Directionality of substrate translocation of the hemolysin A Type I secretion system. *Scientific Reports* 5:12470.

Leopold S, Sawyer S, Whittam T, Tarr P. 2011. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. *BMC Evolutionary Biology* 11:183.

Lescat M, Clermont O, Woerther PL, Glodt J, Dion S, Skurnik D, Djossou F, Dupont C, Perroz G, Picard B, et al. 2013. Commensal *Escherichia coli* strains in Guiana reveal a high genetic diversity with host-dependant population structure. *Environmental Microbiology Reports* 5:49-57.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13:2178-2189.

Liu R, Ochman H. 2007. Stepwise formation of the bacterial flagellar system. *Proceedings of the National Academy of Sciences* 104:7116-7121.

Liu S, Jin D, Lan R, Wang Y, Meng Q, Dai H, Lu S, Hu S, Xu J. 2015. *Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. *International Journal of Systematic and Evolutionary Microbiology* 65:2130-2134.

Lobkovsky A, Wolf Y, Koonin E. 2014. Estimation of prokaryotic supergenome size and composition from gene frequency distributions. *BMC Genomics* 15:S14.

Lobkovsky AE, Wolf YI, Koonin EV. 2013. Gene frequency distributions reject a neutral model of genome evolution. *Genome Biology and Evolution* 5:233-242.

Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* 113:5970-5975.

Lorén JG, Farfán M, Fusté MC. 2014. Molecular phylogenetics and temporal diversification in the genus *Aeromonas* based on the sequences of five housekeeping genes. *PLoS ONE* 9:e88805.

Lukjancenko O, Wassenaar TM, Ussery DW. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology* 60:708-720.

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. *Proceedings of the National Academy of Sciences* 108:7200-7205.

Martin AP, Costello EK, Meyer AF, Nemergut DR, Schmidt SK. 2004. The rate and pattern of cladogenesis in microbes. *Evolution* 58:946-955.

Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research* 40:e6.

Maurelli AT. 2007. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiology Letters* 267:1-8.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genetics* 5:e1000565.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nature Review Microbiology* 10:13-26.

Morlon H, Kemps BD, Plotkin JB, Brisson D. 2012. Explosive radiation of a bacterial species group. *Evolution* 66:2577-2586.

Morlon H, Potts MD, Plotkin JB. 2010. Inferring the dynamics of diversification: a coalescent approach. *PLoS Biology* 8:e1000493.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3:418-426.

Ochman H, Wilson AC. 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *Journal of Molecular Evolution* 26:74-86.

Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, et al. 2009. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proceedings of the National Academy of Sciences* 106:17939-17944.

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, et al. 2014. The SEED and the Rapid

Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* 42:D206-D214.

Parales RE, Ingraham JL. 2010. The surprising *Rut* pathway: an unexpected way to derive nitrogen from pyrimidines. *Journal of Bacteriology* 192:4086-4088.

Park J, Zhang Y, Buboltz AM, Zhang X, Schuster SC, Ahuja U, Liu M, Miller JF, Sebaihia M, Bentley SD, et al. 2012. Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics* 13:1-17.

Peng J, Yang J, Jin Q. 2009. The molecular evolutionary history of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Infection, Genetics and Evolution* 9:147-152.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2: approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.

Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology* 12:1-19.

Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proceedings of the National Academy of Sciences* 97:10567-10572.

Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology* 190:6881-6893.

Retchless AC, Lawrence JG. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences* 107:11453-11458.

Retchless AC, Lawrence JG. 2007. Temporal fragmentation of speciation in bacteria. *Science* 317:1093-1096.

Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences* 106:19126-19131.

Sahl JW, Caporaso JG, Rasko DA, Keim P. 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ* 2:e332.

Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B, Breiman RF, Gilmour M, Nataro JP, et al. 2015. Defining the phylogenomics of *Shigella* species: a pathway to diagnostics. *Journal of Clinical Microbiology* 53:951-960.

Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology* 70:1999-2012.

Sauer MM, Jakob RP, Eras J, Baday S, Eris D, Navarra G, Berneche S, Ernst B, Maier T, Glockshuber R. 2016. Catch-bond mechanism of the bacterial adhesin *FimH*. *Nature Communications* 7:10783.

Shapiro BJ, David LA, Friedman J, Alm EJ. 2009. Looking for Darwin's footprints in the microbial world. *Trends in Microbiology* 17:196-204.

Sims GE, Kim S-H. 2011. Whole-genome phylogeny of *Escherichia coli*/*Shigella* group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences* 108:8329-8334.

Skippington E, Ragan MA. 2012. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli–Shigella* genetic exchange communities. *Open Biology* 2:120112.

Smith JM. 1999. The detection and measurement of recombination from sequence data. *Genetics* 153:1021-1027.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nature Review Genetics* 16:472-482.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences* 108:6187-6192.

Takeuchi N, Cordero OX, Koonin EV, Kaneko K. 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biology* 13:1-11.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10:512-526.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* 30: 2725-2729..

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. *Nature Review Microbiology* 8:207-217.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102:13950-13955.

Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* 11:472-477.

Tibayrenc M, Ayala FJ. 2012. Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proceedings of the National Academy of Sciences* 109:E3305–E3313.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics* 5:e1000344.

Treangen TJ, Rocha EP. 2011a. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics* 7: e1001284.

Treangen TJ, Rocha EPC. 2011b. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics* 7:e1001284.

Turrientes M-C, González-Alba J-M, del Campo R, Baquero M-R, Cantón R, Baquero F, Galán JC. 2014. Recombination blurs phylogenetic groups routine assignment in *Escherichia coli*: setting the record straight. *PLoS ONE* 9:e105395.

Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG. 2013. Comparison of the *Legionella pneumophilapopulation* structure as determined by sequence-based typing and whole genome sequencing. *BMC Microbiology* 13:1-19.

Unno T, Han D, Jang J, Lee S-N, Kim JH, Ko G, Kim BG, Ahn J-H, Kanaly RA, Sadowsky MJ, et al. 2010. High diversity and abundance of antibiotic-resistant *Escherichia coli* isolated from humans and farm animal hosts in Jeonnam Province, South Korea. *Science of The Total Environment* 408:3499-3506.

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. *Nucleic Acids Research* 43:6761-6771.

Villarreal-Chiu JF, Quinn JP, McGrath JW. 2012. The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. *Frontiers in Microbiology* 3:2012.00019.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annual Review of Genetics* 47:97-120.

Vos M, Didelot X. 2008. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 3:199-208.

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009. Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbiology* 75:6534-6544.

Williams D, Gogarten JP, Papke RT. 2012. Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biology and Evolution* 4:1223-1244.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology* 60:1136-1151.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains *Archaea*, *Bacteria*, and *Eucarya*. *Proceedings of the National Academy of Sciences* 87:4576-4579.

Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, Zhu Y, Yu J, Jin Q. 2007. Revisiting the molecular evolutionary history of *Shigella* spp. *Journal of Molecular Evolution* 64:71-79.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24:1586-1591.

Zhang Y, Lin K. 2012. A phylogenomic analysis of *Escherichia coli* / *Shigella* group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evolutionary Biology* 12:1-12.

Zhang Y, Sievert SM. 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in *Epsilonproteobacteria*. *Frontiers in Microbiology* 5:2014.00110.

Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, Wu J, Xiao J. 2014. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* 30:1297-1299.

국문초록

세균의 진화는 다양한 메커니즘을 통해 생성되는 유전체의 다양성을 토대로 이루어진다. 유전체 수준에서 진화가 일어나는 과정을 면밀하게 알아내기 위해서는 이종간의 비교연구보다 종-내에서 벌어지는 유전체 변이 양상을 관찰하는 것이 효과적이다. 대장균은 다양한 생태학적 지위를 가진 종인 동시에 많은 유전체 데이터가 확보되어있기 때문에 세균 유전체의 종-내 다양성 및 진화 과정을 연구하기 위한 이상적인 생물종이다. 본 연구에서는 대장균 종을 대상으로 종-내 유전체 다양성을 여러가지 측면에서 살펴보았고 유전체 변이가 발생하는 과정에서 어떠한 메커니즘이 작용했는지 살펴보았다. 추가적으로 대장균이라는 종이 현재의 계통을 가지게 되기까지의 진화사적 맥락을 근접 종과의 관계 분석을 통해 살펴보았다.

대장균이 보유한 유전체 다양성을 여러가지 관점에서 파악하기 위해 해당 종이 가진 유전자 풀의 다양성, 각 유전자 별로 보유한 염기서열의 다양성, 그리고 유전체 수준에서의 구조적 변이 정도를 각각 분석하였다. 또한 개체 간에 보여지는 유전체의 변이가 어떠한 개체군 수준에서의 구조를 형성하는지 파악하기 위해 계통학적 분석과 집단유전학적 분석을 수행하였다. 분석 결과 대장균의 pan-genome 은 열린 속성을 지니고 있어 이 세균 종의 유전자 풀을 한정할 수 없다는 것을 시사하였다. 이러한 열린 속성은 종 전체의 pan-genome 보다 아종 수준의 계통 별로 측정된 pan-genome 이 더욱 강하게 나타났다. 종-내에서 나타나는 염기서열의 다양성은 core 유전자 전체에 걸쳐 평균 염기서열 다양성

1.3%를 중심으로 한 정규분포 형태로 나타났으며 이는 다수의 세균 종들과 비교하여 일반적인 수준이었다. 유전체를 이루는 유전자 구성에 있어 개체 간의 변이가 큼에도 불구하고 유전자 순서와 SNP 간의 연관분석에 따르면 대장균 종 안에서 염색체 전체에 걸친 core 유전체의 뼈대 구조는 일정하게 나타났다. 종합적으로 유전체 다양성의 분석 결과 중요한 결론은 대장균 안에서 유전체의 구성이 다양화 되는 과정과 그 속에 담긴 염기서열이 다양화 되는 과정이 서로 독립적인 속도로 진행된다는 것과 대장균의 개체군이 아종 수준에서 계통학적으로 뚜렷한 계통들로 나누어진다는 것이다.

대장균의 개체들이 보이는 유전체 다양성이 어떤 기작들을 통해 발생하고 조절되었는지 설명하기 위하여 돌연변이, 상동 유전자에서 일어나는 재조합, 외부 종에서 유래한 수평적 유전자 도입, 자연 선택에 의한 변이 억제 현상을 살펴보았다. 세균의 유전체 다형성은 무성생식을 통해 클로널하게 축적되는 돌연변이와 염색체 재배열에 의해 발생할 뿐만 아니라 동종 혹은 이종 사이에서 벌어지는 다양한 방식의 유전물질 교환에 의한 유성생식적인 변이 확산과 수평적 유전자 도입에 따라 발생하기도 한다. 대장균의 유전체 진화에서 유전자 재조합에 의해 유입된 염기서열은 전체 core 유전체 중 최소 0.78%, 최대 14%, 평균 4.1%를 차지하는 것으로 나타났고 돌연변이에 의한 염기서열 변이에 대비하여 재조합에 의한 염기서열 변이의 상대적 기여도는 0.6 에서 0.8 사이의 값을 가진 계통이 가장 많았다. 계통수에서의 R/theta 값 분포에 따르면 root 에 가까울수록 높은 값을 가진다. 따라서 대장균의 진화 과정에서 재조합의 영향은 점차 감소하여 현재의 0.6 - 0.8 값을 가지게

된 것으로 추정된다. Pan-genome 의 30% 이상을 차지하는 singleton 유전자들은 대장균에서 외부 종 유전자 도입이 활발한 증거로 추정되었기 때문에 그 유전자들의 기원을 밝혀보았다. 절반에 가까운 12,026 개 유전자에 대해서는 어떤 다른 유전체에서도 유사한 유전자를 찾을 수 없었다. 다수 유전자에 대해 아미노산 서열 유사도가 매우 높은 다른 종의 유전자를 찾을 수 있어 해당 종들과 해당 유전자를 공유한 것으로 추정된다. 유전자를 공유한 종은 대장균이 속한 *Enterobacteriales* 목(目)에서 가장 흔하게 나타났지만 분류학적으로 다른 문(門)에 속하는 동떨어진 세균 종과 고세균 종까지 유전자 공유 현상이 발견되었다. 염기서열 변이의 축적은 자연선택에 의해 억제 혹은 촉진될 수 있다. 대장균의 유전자들에 대해 염기서열의 dN/dS 비율을 분석한 결과 일반적으로 유전자별 평균 dN/dS 비율이 1 보다 0 에 가까운 값을 가져 대부분의 유전자에서 아미노산 서열의 변화를 억제하는 방향으로 자연선택이 작용하는 것으로 보였다. 종 내에서 빈도가 높은 유전자들 즉 core 유전자일수록 그러한 경향은 더욱 두드러졌다. 유전자 98 개는 1 보다 큰 dN/dS 값을 보여 아미노산 변이를 선호하는 자연선택이 작용한다는 증거를 보였는데, 그러한 유전자는 대부분 기능이 알려지지 않았으며 core 유전자는 단 2 개만 포함되었다. 일부 기능을 밝힐 수 있는 유전자 중에는 Transposase 단백질의 유전자가 가장 많이 포함되어 있었다. Core 유전자 중에는 편모 (flagella) 생합성 유전자와 기능을 알 수 없는 유전자가 이에 속했다. 종합적으로, 현재의 대장균 pan-genome 에는 다양한 세균 종에서 유래한 유전자와 현재로서 어떤 종으로부터 유래하였는지 알 수 없는 많은 수의 유전자가

존재한다. 염기서열 수준에서 보면 현재는 과거보다 더욱 클로널한 유전체 진화가 이루어지고 있으며 전체 세균의 무성-유성생식적 유전 방식의 스펙트럼의 중간지대에 위치하는 진화 방식을 보였다. 또한 소수의 유전자를 제외하면 자연 선택에 의하여 아미노산 변이가 억제되는 자연선택의 경향이 강하여 염기서열 다양화에 영향을 준 것으로 보인다.

대장균 유전체의 진화 과정을 이해하기 위해 마지막으로 계통분류학 관점에서 대장균 종이 탄생하게된 과정을 살펴보고 인접 종과의 관계를 정리해보았다. 대장균이 속한 *Enterobacteriaceae* 과(科)에 속하는 21 개의 주요 속(屬)의 진화 관계를 809 개의 유전자를 사용하여 분석한 결과 *Enterobacteriaceae* 의 다양한 종은 방사형(radiation)의 분기를 통해 등장한 것으로 보이며 근저에서 빈번한 염기서열 상사성(homoplasy)이 검출되는 특징을 보였다. 기존에 측정된 대장균과 *Salmonella enterica* serovar Typhimurium 의 공통조상에 대한 연대 결정 결과가 1 억 2 천만 - 1 억 6 천만 년 전으로 나타났던 것을 기준점으로 사용하여 대장균과 가장 인접한 근연종들을 포함한 분자시계(molecular clock) 분석을 수행한 결과 대장균이 분화된 시점은 1660 만 - 1770 만 년 전으로 추정되었다. 대장균이 속한 *Escherichia* 속에서 대장균과 기타 계통들의 진화적 분기 순서를 결정하기 위해 Bayesian species tree 분석을 해보았다. 유전체 데이터를 사용하여 다수의 유전자를 분석하였음에도 유전자 별로 기록된 분기 순서가 상이하여 정확한 분기 순서를 결정할 수 없었다. 이 외에도 계통분류학적으로 모호하게 정의된 인체 병원성 세균 4 종을 포함하는 *Shigella* 속과 대장균 종의 진화

관계를 명확히 정의하고자 *Shigella* 병원성의 유전적 원인으로 여겨지는 독성 플라스미드의 유전자들이 가지는 계통수와 그들의 염색체 유전자들이 가지는 계통수를 비교 분석하였다. 플라스미드 획득 경로를 설명하기 위해서는 *Shigella* 4 회 이상 독립적인 플라스미드 획득이 있었던 것으로 설정해야 두 계통수에 위배되지 않는 설명이 가능했다. 종합적으로, 유전체 데이터를 통해 많은 정보를 분석에 사용하더라도 대장균과 인접 종들의 진화적 관계를 명확히 제시할 수는 없으며, 그 이유로 실제 진화적 분기가 짧은 시간 안에 두갈래로 종분화되는 방식으로 일어나지 않았을 가능성을 들 수 있다.

세균에서 종의 유전학적 개념은 간단하게 정의하기 어렵다. 무성-유성생식 사이에서 종마다 위치가 상이하며 이와 별개로 종에 따라 pan-genome 이 상대적으로 빠른 변화를 보이거나 느린 변화를 보이는 차이를 가진다. 그러한 진화 방식의 차이에 따라 종 분화의 기작도 상이하게 나타나는 것으로 보인다. 이번 연구에 따르면 대장균은 과거 비교적 유성생식적으로 진화하는 개체군에서 점차적으로 비교적 무성생식 쪽으로 전환되며 몇 개의 계통으로 나뉘어진 것으로 추정된다. 이러한 집단유전학적 특성 전환은 대장균에서 뚜렷한 계통분류학적 구조가 나타난 배경으로 사료된다. 무성 생식에 가까울수록 core 유전체의 염기서열 다양화가 지연됨에 따라 상대적으로 염기서열 진화에 비해 유전체 변화의 비중이 커지는 것으로 보인다.

주요어: 대장균, 세균, 유전체학, 진화, 종, 계통학, 범유전체