VARIATION IN CORE AND ACCESSORY PARTS OF GENOME OF *ESCHERICHIA COLI* ISOLATED FROM SOIL FROM RIPARIAN AREAS IN NEW YORK STATE

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Oleksandr Maistrenko

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Genomics and Bioinformatics

March 2016

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Variation in core and accessory parts of genome of *Escherichia coli* isolated from soil from riparian areas in New York State

By

Oleksandr Maistrenko

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Peter Bergholz
Chair

Dr. Changhui Yan

Dr. Phillip McClean

Approved:

04/08/2016	Dr. Phillip McClean
Date	Department Chair

ABSTRACT

Escherichia coli is commensal bacteria and is a symbiont of the digestive system of vertebrates. Due to frequent deposition of *E. coli* into extrahost habitats (soil, water), approximately half of its population exists as free living organisms. It is unclear what genome-wide variation stands behind adaptation for extrahost habitat. This thesis applies a genome-wide association study approach to find genetic variation in core and accessory parts of genome of *E. coli* that is associated with 1) forest or agricultural field soil habitats and 2) with survival phenotype in soil microcosm. Gene composition analysis suggests that pan-genome of environmental *E. coli* is unlimited. Core and accessory genome contained variation associated with survival phenotype and with forest or field habitat.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Peter Bergholz for advising and for giving me an opportunity to work in his laboratory. I would like to acknowledge Kaycie Schmidt and Julie Sherwood for helping me with the work in the laboratory. I would also like to thank my supervisory committee, Dr. Phillip McClean and Dr. Changhui Yan, for guidance through my master's studies.

I would like to acknowledge the research team in Food Safety Laboratory at Cornell University for soil sampling and isolation of *E. coli* from soil: Laura Strawn, Steven Warchoki, Gina Ryan, Courtenay Simmons and Jihun Kang, Dr. Daniel Buckley, Cheryl Andam, Jesse Noar, Nicole Birrer.

I'd like acknowledge the funding sources that allowed me to perform this research and for giving me an opportunity to study at North Dakota State University: Federal Formula Funding (Hatch-Act), ND-EPSCoR, and Fulbright-STEP scholarship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. LITERATURE REVIEW

Microbial species and flexibility of microbial genomes

Historically the definition of microbial species and microbial systematics was based on cultural characteristics of organisms (Puntoni 1952; Sharma et al. 2015). Early approaches largely underestimated microbial diversity because they mostly covered only cultivable bacteria (Staley and Konopka 1985). Precise delineation of the new microbial species according to the current edition of *International Code of Nomenclature of Bacteria* should be performed using several types of markers: Amplified Fragment Length Polymorphisms, Random Amplified Polymorphic DNA, Repetitive Element Palindromic PCR, Pulsed-field Gel Electrophoresis of gene clusters (ribotyping of rRNA operons), typing of individual genes (Amplified rDNA Restriction Analysis of 16S rDNA) and intergenic 16S–23S rDNA spacer regions, DNA-DNA-hybridization, Multilocus Sequence Typing, as well as several physical methods: infrared spectroscopy and several types of spectrometry (Stackebrandt et al. 2002). Complete documentation of microbial species relies on the isolation of putative species into laboratory culture (Yarza et al. 2014).

Species concept itself is a subject of debates. At present, there are more than 22 species concepts (Mayden 1997; Naomi 2011) that take into account different aspects of organisms' biology to define species. Current concept of bacterial species relies on three postulates: members of species are monophyletic group, and they exhibit genotypic and phenotypic coherencies (Rossello-Mora and Amann 2015). Sometimes, microbes are classified into species based on average nucleotide identity. If average nucleotide identity between strains is more than or equal to 95%, they are considered the same species (Konstantinidis et al. 2006). Cutoff of 95% can be flexible for different microbial groups.

Biological species concept, which is widely used for Eukaryotes, relies on biparental sexual reproduction mode (Mayden 1997). Sexual recombination in microbes occurs only very rarely, but gene conversion and horizontal gene transfer (HGT) are much more frequent and occurs through conjugation, cell fusion, transduction, and transformation (Soucy et al. 2015). The above processes lead to breakdown of linkage disequilibrium that would arise in clonal populations (Bobay et al. 2015).

Microbial genome is an open system which experiences constant loss and acquisition of genes (Puigbo et al. 2014). The plasticity of microbial genomes is ensured by ability to feed with DNA and by mobile genetic elements: transposons, bacteriophages, plasmid exchange (Darmon and Leach 2014). HGT occurs more frequently within species and between closely related subgroups of microbes within the same species. HGT is limited by several factors: 1) bacteriophages often have restricted host range due to specialization in adherence, invasion and replication genes (Chibani-Chennoufi et al. 2004); 2) plasmids preservation is limited due to plasmids incompatibility (Velappan et al. 2007); 3) CRISPR/Cas system can provide immunity to foreign DNA infection (Garneau et al. 2010); 4) other restriction systems enable the utilization of foreign DNA which frequently acts as a nutrient source rather than as a donor of new genetic information (Finkel and Kolter 2001). For these reasons, it has been proposed that recombination rate and gene content may be used in some cases as markers of microbial species or subspecies groups (Konstantinidis et al. 2006).

Due to high plasticity of gene content each strain of bacterial species contains genes that aren't present in any other strain of the same species. Pan-genome is a set of all unique genes that can be ever encountered in genomes of taxonomic group of organisms (Medini et al. 2005; Tettelin et al. 2005). In general, size of genome and number of genes are taxon-specific traits. However,

on average up to 50% of genes in a microbial genome may be absent in every genome of the same species (Tettelin et al. 2008). These genes are called accessory genes. Genes that are present in every genome of a taxonomic group of an organism compose the core genome. The core genome primarily consists of housekeeping genes involved in replication of DNA, transcription and translation machinery, cell envelope maintenance, energy metabolism, regulatory functions, and transport and binding proteins (Medini et al. 2005; Vernikos et al. 2015). Approximately, one-third of the shared (core) genes encode hypothetical or unknown function proteins. The core genes together with all other accessory genes that are ever encountered in species compose the pan-genome (Tettelin et al. 2008).

Microbial population is a group of coexisting individuals which have arisen from a single or small set of common ancestors. Individuals from the same population are highly clustered on the genotypic and phenotypic levels, meaning that the variance between populations is much greater than the variance within populations (Polz et al. 2006). Due to the large size, temporal instability of the environment and competition for resources one population of the same microbial species can diverge into several ecotypes that occupy different ecological niches. An ecotype is a group of individuals which were selected to occupy a similar ecological niche within the community or ecosystem (Cordero and Polz 2014a). Ecotypes and populations of one species can have somewhat isolated pan-genomes (Boon et al. 2014; Reno et al. 2009). For example, in *Sulfolobus islandicus* spatial isolation of populations led to local adaptation and divergence into ecotypes with somewhat isolated pan-genomes among caldera basins (Reno et al. 2009).

Microevolution in bacteria. Evolutionary models

Several models have been developed to describe how bacterial ecotypes evolve and speciation occurs depending on the contribution for adaptation of variation in core and accessory

parts of the genome. According to stable ecotype model individuals with adaptive mutations in the either core or accessory genome in the population of the ecotype outcompete all other less adapted specimens of the same ecotype. Consequences of such events are periodic purging of ecotypes' genetic diversity, while diversity in other ecotypes remains intact (Cohan 2001). This model assumes that recombination is not high enough to unlink adaptive and neutral loci in the genome. Consequently, the rest of the genome "hitchhikes" with adaptive variant. This model was inspired by experimental evolution studies. Due to high rates of recombination and HGT it possible that this model limitedly explains evolution of microorganisms in nature (Shapiro and Polz 2015).

Recursive niche invasion model postulates that adaptive trait of ecotype can be acquired by HGT and lost rapidly. Ecotype specialization is reversible on short scale and depends only on presence/absence of genes which allow occupation of an ecological niche (Godreuil et al. 2005).

Nano-niche model suggests that subgroups of same ecotype become incompletely diverged in response to conditions of complex habitat (Cohan 2005). Selection and spatial isolation can either lead to complete ecotype divergence or extinction of subgroups of ecotype. Adaptation is achieved by different genetic variants located in either core or accessory genome. The core and accessory genome interact less with each other than in stable ecotype model because of fine grained changes in fitness.

Considering that 1) different microorganisms have different rate of HGT and 2) genes that are responsible for phenotypic traits can belong to the core genome and/or to the accessory genome, evolution of respective ecotypes (associated with given phenotypic trait) can resemble different evolutionary models (Shapiro and Polz 2015).

Population genetics of *E. coli*

Escherichia coli is widely known Gram-negative, non-sporulating facultative anaerobe microorganism which in natural habitat is symbiont of the digestive system of vertebrates (Gordon and Cowling 2003). *E. coli* has a complex life cycle which includes frequent deposition into secondary habitats, including surface soils and water (Blount 2015; Winfield and Groisman 2003). Total wild population size is about 10^{20} cells (Whitman et al. 1998), approximately half of which at any moment of time persists outside host organisms (Savageau 1983). Type of interactions between host organism and *E. coli* can vary across the entire spectrum of host-ectosymbiont interactions from mutualism to commensalism and to parasitism (Croxen et al. 2013).

Divergence of the *E. coli* from the closest relative *Salmonella* occurred approximately 120-160 Myr ago (Ochman and Wilson 1987). Currently, *E. coli* is subdivided into seven clades (*sensu stricto E. coli*) that are called phylotypes A, B1, B2, C, D, E, and F, and four *Escherichia* cryptic clades I-IV. Phylogroups are most commonly distinguished based on sequence types in genes *arpA, chuA, yjaA* and TspE4.C2 (Clermont et al. 2000; Clermont et al. 2013), but some of these groups are only distinguished with approximately 79-95% accuracy using this method. Average genome of *sensu stricto E. coli* is composed of 4,721 genes; core and pan-genome are composed of approx. 2,000 and 18,000 genes, respectively (Hendrickson 2009; Tenaillon et al. 2010; Touchon et al. 2009). Population of *E. coli* is viewed as a set of clonal lines with frequent recombination/"hybridization" events among some lineages (Wirth et al. 2006). Recombination is more common in B1, B2 and D phylotypes, and hybrid lineages exist, including, AxB1 and AxBxD. Further studies showed that phylotypes A and B1, B2, and E possibly are at early stages of speciation because level of recombination within these phylotypes is higher than between each other (Didelot et al. 2012).

Microbial community assembly. Life cycle of *E. coli*

When *E. coli*, or any microorganism, enters new habitat it becomes affected by forces that assemble microbial communities, including abiotic and biotic selective factors (Sikorski 2015). Action of community assembly forces on the microorganism are likely to cause selective sweeps of initial genetic variation of invader and acquisition of new variation. It is important for this reason to understand what process and what factors from microbial community side are prerequisites for the ecotype divergence of invading microorganism (Sikorski 2015).

A microbial community is a group of microorganisms that co-occur in space in time. Communities are assembled with four community assembly forces: selection, ecological drift, dispersal/admixture and speciation (Martiny et al. 2006; Vellend 2010). Community assembly forces are somewhat different from evolutionary forces (mutations/HGT, selection, drift, dispersal) because they act on the group of species rather than on the population of single species.

Soil microbial community is selected by physical factors of soil: content of inorganic and organic substances, pH, temperature, moisture (Lozupone and Knight 2007). Biotic factors such as type of vegetation also have effect on structure and assemblage of the microbial community. Previous studies showed that structure of soil microbial community responds to deforestation (Crowther et al. 2014). Magnitude of response depended on the soil texture: the microbial community of sandy soil experienced stronger shifts than communities from types of soil with finer texture. Structure of microbial community also varies depending on proximity from plant roots (Mendes et al. 2013). Effect of physical properties of soil (e. g. pH, temperature, C:N ratio) can have greater effect on the structure of the microbial community than biotic factors such as type of land cover or land usage (Kuramae et al. 2012).

Structure of the microbial community is determined by dispersal processes that are based on the ability/chance of microorganism to migrate to other communities and invade them (Mallon et al. 2015). Size and content of microbial community is also shaped by stochastic process – drift. For example, in the study Stegen et al (2013) up to 25% of microbial community composition shifts were attributed to stochastic processes (Stegen et al. 2013).

The factors that shape the microbial community structure (selection, drift, dispersion, speciation) act in any combination together (Nemergut et al. 2013). The effect of this forces can result in different preservation patterns of population of any given microbial species. Prolonged co-occurrence of several species of microorganisms together with prolonged influence of drift, selection and dispersal process can lead to the ecotype divergence and/or speciation (Sikorski 2015).

E. coli is presented in the ecosystem as an ectosymbiont (predominantly as a commensal or mutualist) of vertebrate animals and as a free-living microorganisms of soil and water. Large portion of *E. coli* population in extrahost environment is composed of naturalized previously symbiotic specimens adapted to persist and proliferate outside animal host (de los Angeles Dublan et al. 2014; Texier et al. 2008; Walk et al. 2009a). Characterization of population of *E. coli* isolated from same sampling site (Dairy Alpine Grassland Soils) by sequence of *uidA* gene (encodes the β-D-glucuronidase protein) showed that naturalized and fecal isolates compose two distinct clades (subpopulations) (Texier et al. 2008). Significant part of *E. coli* population consists of recently deposited individuals that can survive in extrahost environment only for a short time from 1-2 days up to several months (Berthe et al. 2013) before encountering new host, evolutionary and/or demographic and/or genetic rescue or extirpation otherwise. Evolutionary rescue happens when acquired genetic variants (*via de novo* mutations or HGT) facilitate appearance of adaptive

phenotype which in turn allows a population to recover from decline initiated by environmental conditions (Gonzalez et al. 2013). Demographic and/or genetic rescues occur due to constant supplying of environmental population of *E. coli* with newly deposited specimens into the extrahost habitat (Carlson et al. 2014). If rescue occurs for commensal strain of *E. coli* it becomes a part of naturalized subpopulation.

Ecological niche of *E. coli*.

There is a significant niche divergence between phylogroups of *E. coli* (Table 1). For example, phylotypes are unequally distributed across different hosts. *E. coli* of phylotype A is dominant in human (40.5%). B2, B1 and D are less frequent in human: 25.5%, 17%, and 17% respectively. In other animals, abundance ratios of phylotypes are quite different with B1 at 41%, A at 21%, B2 at 21%, and D at 16% (Tenaillon et al. 2010). In extrahost environments, phylotypes B1 and D are more prevalent than any other (Bergholz et al. 2011; Orsi et al. 2007; Ratajczak et al. 2010), whereas phylotype A is found rarely in soil habitats but more frequently in water. *E. coli* is frequently found to proliferate and/or persist on plants (Brandl 2006). It was shown that phylotypes B1 and D had strong association with plants, A and E also showed association with plants but to lesser extent (Meric et al. 2013). Recent study showed that there is temporal variation of the phylotype prevalence: phylotype A is more prevalent in low temperature part of season and B in high temperature part of season (Jang et al. 2014).

Table 1. Association of *E. coli* with different host and habitats.

Phylotype	Water, %	Soil, %	Human, %	Other animals, %	Plants
A	30.6	3.4	40.5	21	Low association
B1	38.8	39.0	17	41	High
B2	4.1	20.0	25.5	21	Low
D	26.5	37.6	17	16	High
References	Orsi et al., 2007	Bergholz et al., 2011	Tenaillon et al., 2010		Meric et al., 2013

One of key questions in microbial ecology is what genome-wide variation stands behind the different ecological properties (e. g. association with host, survival in the extrahost habitat, etc.) of ecotypes of bacteria. Various researches address this question by studying what variation in the core and accessory parts of *E. coli* genome display association with host and extrahost habitats. It was suggested that genes from the core genome that display variation associated with different persistence phenotypes in extrahost environment tend to belong to stress response and metabolic flux functional groups (van Elsas et al. 2011a). In *E. coli* importance of variation was repeatedly showed for *rpoS* core gene, the general stress response sigma factor, for extrahost persistence in *E. coli* (Rozen and Belkin 2001). Poor persistors among *E. coli* O157:H7 in soil (<160 days) tended to carry SNPs, insertion, and deletions in *rpoS* gene compared to long term survivors (>200 days) (van Hoek et al. 2013). However, several other studies have questioned presence of disruptive variation in *rpoS* in environmental isolates. It was suggested that variation in this gene is rather reflection of "source-sink" dynamics during laboratory cultivation of environmental isolates of *E. coli* (Ihssen et al. 2007; Snyder et al. 2012).

The accessory genome appears to be important for structuring of microbial population across spatial and temporal dimensions of the ecological niche. Study on *Vibrio* showed that,

among accessory genes, prophage genes, hypothetical, and/or unstructured proteins were associated with temporal and spatial structuring of population (Dutilh et al. 2014b). Recent study on *E. coli* of phylotype B1 by using PCR-based fingerprinting of accessory genes to find accessory variants associated with fecal and environmental strains of *E. coli* showed that toxin-antitoxin system accessory genes were more abundant among fecal strains, whereas genes involved in iron acquisition, complement resistance/surface exclusion, and biofilm formation were more abundant among environmental strains (Tymensen et al. 2015).

Yet, many studies on microbes aiming to address the question how variation in the core and accessory genome is associated with phenotype and ecological properties of microbial species use only limited quantity of genomic variation obtained, for example, by gene typing (Read and Massey 2014). Alternative approach is to use genome-wide association studies on whole-genome sequences of bacteria. Genome-wide association studies (GWAS) are applied to understand the maintenance and spread of variation for extrahost survival, habitat association, virulence, host preference and antibiotic resistance (Chen and Shapiro 2015; Salipante et al. 2015; Sheppard et al. 2013). Main advantage of GWAS is that it allows to account for most of the variation available in the genome (Korte and Farlow 2013) compared to multi-gene typing methods. Genotype – phenotype associations are discovered with comparison of means (ANOVA, Mann–Whitney U, Kruskal–Wallis tests etc); correlational analysis (Pearson's chi-squared test, Kendall tau rank correlation coefficient and Spearman's rho correlation) and machine learning (random forest and support vector machines) (Dutilh et al. 2013; Ziegler et al. 2008). Machine learning approaches are becoming increasingly popular for GWAS (Szymczak et al. 2009). Main advantage of the machine learning approach is that it utilizes multiple variables at once and allows prediction of phenotype based on combination of variables which on their own don't have predictive power

(Lunetta et al. 2004). Meanwhile, main disadvantage of the machine learning approaches is the absence of explicit importance measures for interactions between genetic variants (Dutilh et al. 2013).

CHAPTER 2. VARIATION IN CORE AND ACCESSORY PARTS OF GENOME OF *ESCHERICHIA COLI* ASSOCIATED WITH DIFFERENT EXTRAHOST HABITATS: SOIL FROM FOREST AND AGRICULTURAL FIELD

Introduction

Escherichia coli is facultative anaerobe microorganism and is a symbiont of the digestive system of vertebrates (Gordon and Cowling 2003). *E. coli* is frequently deposited into secondary habitats, including surface soils and water (Blount 2015; Winfield and Groisman 2003). Approximately half of *E. coli* population persists outside host organisms (Savageau 1983). *E. coli* is subdivided into 7 clades (*sensu stricto E. coli*) that are called phylotypes A, B1, B2, C, D, E, and F, and four *Escherichia* cryptic clades I-IV (Clermont et al. 2000; Clermont et al. 2013). Population of *E. coli* is viewed as a set of clonal lines with frequent recombination/"hybridization" events among some lineages (Wirth et al. 2006). An average genome of *sensu stricto E. coli* is composed of 4,700 genes; the core and pan-genome are composed of approximately 2,000 and 18,000 genes, respectively (Hendrickson 2009; Tenaillon et al. 2010; Touchon et al. 2009). Horizontal gene transfer (HGT) plays an essential role in evolution of *E. coli* genome. Previous studies showed that the pan-genome of commensal and pathogenic strains of *E. coli* is potentially unlimited (Didelot et al. 2012; Touchon et al. 2009).

Preservation of any microorganism, including *E. coli*, in the extrahost habitat is mediated by action of abiotic and biotic factors of the habitat (van Elsas et al. 2011a). In soil, variation of the abiotic factors (content of inorganic and organic substances, pH, temperature, moisture, and spatial structure/texture of habitat) together with variation of the biotic factors (species composition and land cover) create different habitat complexity (Kovalenko et al. 2012). Land cover and land usage affects structure of soil microbial community. For example, soil microbial

community responds to deforestation (Crowther et al. 2014) and varies depending on proximity from plant roots (Mendes et al. 2013). Other study showed that physical properties of soil (e. g. pH, temperature, C:N ratio) can have greater effect on structure of microbial community than biotic factors such as type of land cover or land usage (Kuramae et al. 2012). Soils from agricultural field and forests have different land cover, organic and inorganic matter content, soil biota, etc. (Paul 2014). In other words, soil from forest and from agricultural field has different habitat complexity.

Genotype of any given species affects capacity to invade and/or persist in soil community (Mallon et al. 2015). It was repeatedly shown that genetic background affects capability of *E. coli* to survive in the extrahost habitat (van Elsas et al. 2011a). It is also possible that genetic variations in core and accessory genome contributes differently to adaptation to specific habitat complexity. If this is true then soil population of *E. coli* may diverge into several ecotypes which are more associated with habitats of forest soil or agricultural field soil.

Three major models were developed to explain how divergence into ecotypes and speciation occurs in bacteria assuming different contribution of the core and accessory genome variation. According to stable ecotype model, individuals with adaptive mutations in either core or accessory genome in population of the ecotype outcompete all other less adapted specimens of the same ecotype. (Cohan 2001). Recursive niche invasion model postulates that adaptive trait of ecotype can be acquired by HGT and lost rapidly. Ecotype specialization is reversible in the short term (Godreuil et al. 2005). Nano-niche model suggests that subgroups of the same ecotype become incompletely diverged in response to conditions of complex habitat (Cohan 2005). Selection and spatial isolation can either lead to complete ecotype divergence or extinction of subgroups of ecotype. Adaptation is achieved by different genetic variants located in either core or accessory genome.

When microbial populations are subdivided by different environmental selection pressures, genome-wide association studies and population genomic analyses can be used to understand the generation, maintenance and spread of habitat specific adaptations (Cordero and Polz 2014b; Dutilh et al. 2014a; Shapiro et al. 2012).

Major aims of this study are:

1) Evaluate shared and unique genes content and determine SNPs/indels variation in isolates of *E. coli* from soil of phylotypes D and B1.
2) Estimate size of core and pan-genome in phylotypes D and B1.
3) Perform genome wide association study between genome variation with soil habitat complexity (forest soil and agricultural field soil) of phylotypes D and B1.
4) Estimate the relative association of core and accessory genome variation with soil from different habitats (forest or agricultural field).

Materials and Methods

Origin of E. coli strains

E. coli strains were isolated from alluvial silty or sandy loam soil that originates from New York State agricultural field (phylotype D: 85 strains; phylotype B1: 43) and forest (phylotype D: 46 strains; phylotype B1: 13) landscapes near Hoosick River and Flint Creek (Table 2, 3).

Table 2. Coordinates of soil sampling sites near Hoosick River (New York State).

Name	Latitude	Longitude
Field 13A	42.91773	-73.6431
Field 19A	42.907819	-73.6502
Field 19B	42.906329	-73.6506
Field 11	42.895178	-73.634
Forest F11	42.894139	-73.6342
Forest F12	42.894141	-73.6356
Field 10	42.89386	-73.6312
Field 12	42.891779	-73.6334
Forest F14	42.921235	-73.5232
Field 14	42.920248	-73.52
Field 15	42.91884	-73.5128
Forest F15	42.919203	-73.5117
Forest F16	42.950931	-73.3876
Field 16	42.950421	-73.3885
Field 17	42.948771	-73.3871
Forest F17	42.94793	-73.3843
Forest F18	42.946315	-73.3777
Field 18	42.944937	-73.3777
Field 13B	42.91925	-73.6419
Forest F13	42.918729	-73.6454

Table 3. Coordinates of soil sampling sites near Flint Creek (New York State).

Name	Latitude	Longitude
Field 1	42.82364	-77.1214
Field 2	42.83044	-77.1362
Field 3	42.81921	-77.0969
Field 4	42.83847	-77.0672
Field 5	42.8419	-77.1049
Field 6	42.85383	-77.1119
Field 7	42.91033	-77.105
Field 8	42.93536	-77.0968
Field 9	42.93824	-77.0968
Forest F1	42.82588	-77.1244
Forest F2	42.83154	-77.1368
Forest F4	42.83941	-77.0661
Forest F5	42.8445	-77.1057
Forest F6	42.85452	-77.109
Forest F7	42.91007	-77.1078
Forest F8	42.93472	-77.0984
Forest F9	42.93922	-77.0982

Soil sampling and isolation of *E. coli* was performed by research team in Food Safety Laboratory at Cornell University: Laura Strawn, Steven Warchoki, Gina Ryan, Courtenay Simmons and Jihun Kang.

DNA extraction

Total genomic DNA of *E. coli* strains was extracted from overnight (14-16 hours, 215 rpm, 37ºC) cultures grown on LB media (5g/l NaCl, 5 g/l yeast extract, 10g/l peptone). Bacterial suspension cultures were harvested by centrifugation 15,000 rpm (16,600 g) for 10 min (Allegra X-30R Centrifuge; rotor model: BECKMAN F2402H, rotor serial number E2527, diameter of rotor: 132 mm). Total genomic DNA from pelleted bacteria was extracted by phenol/chloroform method. Cell suspensions of bacteria in TE were incubated with 1) 25 µl of 40 mg/ml of lysozyme (at 37 ºC for 2 hours); 2) 6 µl of 20 mg/ml RNAase A to lysate (37 ºC for 10 min); 3) 15 µl 20

mg/ml Proteinase K (at 37 ºC for 1 hour). Obtained lysate of bacterial cultures were mixed with 100 µl of 4 M NaCl and 60 µl 1 M NaCl with 10% CTAB and further incubated 65ºC for 10 min. Cleaning of DNA was performed by series of centrifugations with Tris-saturated phenol (pH 7.9) and chlorophorm:isoamyl alcohol (24:1). DNA was precipitated with Na-acetate and isopropanol. Obtained pellet of DNA was washed in 70% ethanol and dried at room temperature DNA and then was dissolved in 50 µl distillated H_2O (Sambrook at al. 1989).

To verify quality and purity of extracted DNA I used spectrophotometer (NanoDrop, Thermo Fisher Scientific, Wilmington, DE). DNA samples were considered of high quality if they contained at least 20ng/µl of DNA, and absorption ratios at 260/230 were in range of 1.8-2.2 and 280/260 were in range of 1.6-1.8. Quantity of double-stranded DNA was measured using fluorimeter (Synergy H1 Hybrid Reader , BioTek, Instruments, Inc, Winooski, VT) with Quanti-iT™ PicoGreen® dsDNA Reagent and Kits (Invitrogen, Carlsbad, CA) according to protocol provided by manufacturer.

DNA-library preparation and sequencing

Concentration of input double strand DNA was adjusted to 0.2ng/µl. Paired-end sequence libraries were generated using a Nextera® XT Sequence Library preparation kit (Illumina, San Diego, CA) and with oligonucleotide barcodes delineating each strain according to best practices protocol. Fragment size of DNA library was assessed using electrophoresis unit (Agilent 2100 Bioanalyzer, Agilent Technologies, Inc, Santa Clara, CA) according to Illumina recommendations. Sequence reads of 100bp length were obtained using an Illumina HiSeq 2500 at Axeq Technologies sequencing facility (World Meridian Venture Center, Seoul, Republic of Korea).

Sequences analysis

I performed quality control of the paired-end reads using FastQC v.0.11.2 (Andrews S. 2010). Per base sequence quality (measured in Phred score) did not drop less than 20 across all reads data. 5'-end of all reads data had significant bias in base content because of presence of adapters. More than 20% of sequences were duplicated in 82% of files, indicating a high level of coverage. Contamination of reads with Nextera transposase sequences was observed in 26% of raw sequence files. Trimmomatic v.0.32 (Bolger et al. 2014) was used to remove Nextera transposase sequence and remnants of adapters at the beginning of reads (the first 13 nucleotides were removed from all reads). Adaptive trimming was also applied that favored longer trimmed reads (MAXINFO parameter was set at 0.3). All reads of length less than 70 bases after trimming were discarded. Quality control after adaptive trimming verified that percentage of raw sequence files that contained >20% duplication levels was reduced from 82% to 54%, adapter sequences were completely removed, per base sequence content became less biased at the 5'-end of reads across all sequence files (all reads across all sequence files had less than 20% difference between A and T, or G and C bases).

I applied Genome Analysis ToolKit v3.3 (McKenna et al. 2010) for SNPs and indels discovery and hard-filtering procedure taking into account the developers' recommendations (DePristo et al. 2011). Reads for these analysis were mapped using Stampy.py v.1.0.21 (Lunter and Goodson 2011) and bwa-mem v.0.6.2 (Li and Durbin 2009) on reference *E. coli* genomes of strains UMN026 and IAI1 for phylotypes D and B, respectively. Genomic variants were also determined by variant calling procedures using cortex_var v.1.0.5.21 (Iqbal et al. 2012). Annotation of SNPs and indels was performed using snpEff v.4.1 (Cingolani et al. 2012). Only

variants with representation in all sequenced genomes and called by both GATK and cortex_var softwares were retained for downstream analysis.

For gene content analysis, *de novo* assemblies of reads were performed with Velvet v.1.1 (Zerbino and Birney 2008) with automatic expected coverage and coverage cutoff determination with k-mers length at 37 (which is near the optimal k-mer length for Velvet assembler (Haridas et al. 2011)). Summary of *de novo* assembly statistics is presented in Fig. 1. Average depth of coverage varied from 16 up to 91 times. N50 –value varied from 64974 up to 436366 bases.

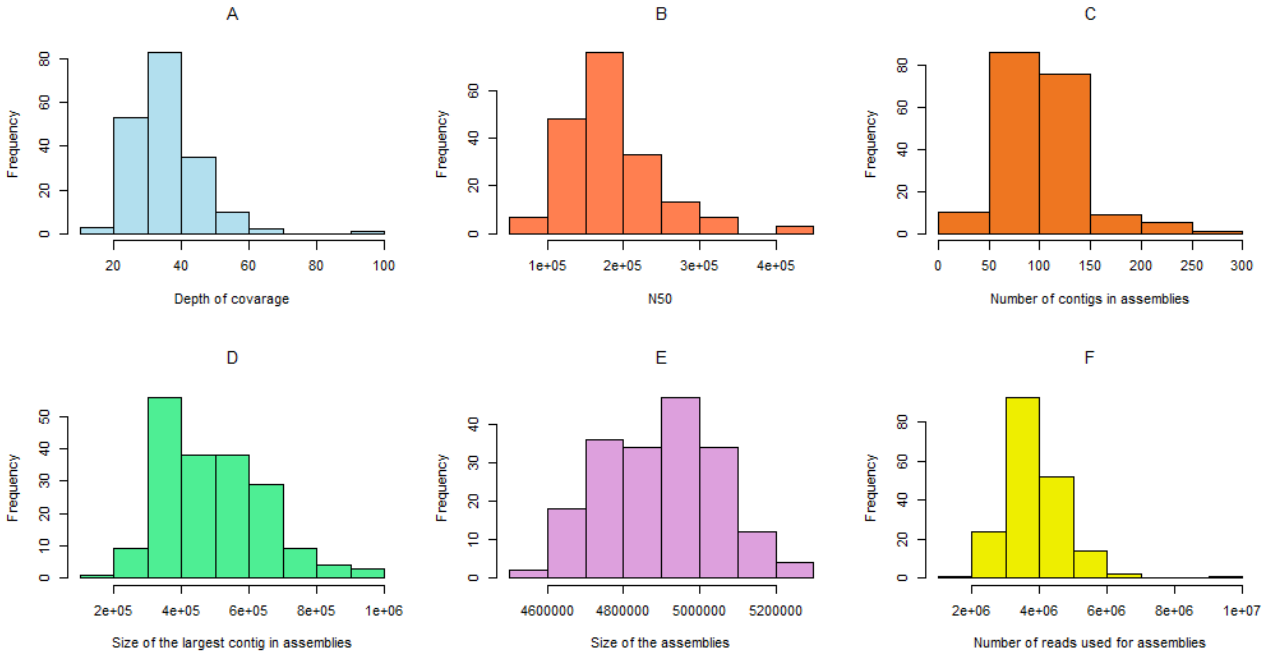

Figure 1. *De novo* assembly statistics of 187 genomes of *E. coli*. Y-axis represents quantity of genomes which had any given value in x-axis. X-axis: A – Depth of coverage; B – N50-value; C – Number of contigs in assembly; D – Size of the largest contig in assemblies; E – Size of the assemblies; F – Number of reads used for assemblies.

Contigs were ordered against reference genomes using Mauve 2.3.1 (Darling et al. 2004). I submitted assemblies to Rapid Annotation using Subsystems Technology (RAST) service (Aziz et al. 2008) to annotate genomic genes, phage genes and mobile elements.

Gene content variation among strains was analyzed by GET_HOMOLOGUES software (Contreras-Moreira and Vinuesa 2013). To identify all orthologous genes within D and B1 phylotypes I used OrthoMCL algorithm (Li et al. 2003). Two genes were considered orthologous (or in other words shared between pair of compared isolates of *E. coli*) if they exhibited ≥80% amino acid sequence identity and ≥75% overlap of the query sequence length in reciprocal BLAST searches between genomes.

Size of the core genome was estimated based on two approaches suggested by Tettelin at al 2005 (Tettelin et al. 2005) and Willenbrock et al 2007 (Willenbrock et al. 2007) using GET_HOMOLOGUES software. Size and content of the pan-genome was estimated according to Tettelin at al 2005. The core and pan-genomes curves were built by ten random input orders of 56 genomes of phylotype B1 and 131 of phylotype D.

Identity matrices were calculated based on the average protein identity between any given pair of genomes. Dendrograms and heatmaps were generated in R 3.2.3 (R Core Team 2015) using packages: dendextend (Galili 2015); ape (Paradis et al. 2004); gplots (Gregory R. Warnes et al., 2015).

To identify and annotate prophages in *E. coli* genome I submitted contigs ordered and concatenated into pseudo-chromosomes to PHAST (Phage Search Tool) service (Zhou et al. 2011).

Genome wide discovery of variants associated with habitat

To find variants in the core and accessory portions of the genome associated with habitat (field or forest), I used a machine learning algorithm – random forest (Breiman 2001). Random forest (RF) is an ensemble of regression or classification trees, each built using a bootstrap subset of the original data. The accuracy of each tree is measured using the remaining test data. In this analysis, the response variable is habitat, and the predictors are SNPs, indels, and accessory genes

(Dutilh et al. 2013; Dutilh et al. 2014b). Mean decrease of accuracy (MDA) was selected as measure of importance of the variants for prediction of strains' origin. Number of variables randomly sampled as candidates at each split was set to mtry=100; number of trees to grow was set at ntree=10,000; ntree=100,000; ntree=200,000 (Boulesteix et al. 2012; Dutilh et al. 2014b). Analysis was performed using randomForest package (Liaw and Wiener 2002) in R 3.2.3 (R Core Team 2015). Matrices containing strain names, type of habitat and binary information about presence or absence of SNPs, indels and accessory genes variants were used as an input for RF. The core and accessory genetic variants with minor allele frequency ≤ 10 % were removed from the analysis. In other words, all variants found in less than in 15 genomes in phylotype D and less than in 6 genomes in phylotype B1 and more than in 116 genomes in phylotype D and more than 51 in phylotype B1 of strains were removed from GWAS. These variants are unlikely to be associated with the habitat because they either are rare variants compared to sample size of strains from specific habitat or they represent difference between studied environmental strains and the reference strain.

Results and discussion

Gene content

I performed estimation of the core and pan-genome size using GET_HOMOLOGUES software (Table 4, Fig. 2 and 3). In phylotype D the size of core genome was 2,003 genes and size of pan-genome was 18,430 genes. Sizes of core and pan-genome in phylotype B1 were 2,761 and 13,374 of genes, respectively (Table 4). Fitted equations (2.1-2.6) were used to extrapolate the core (2.1-2.4) and pan-genome (2.5-2.6) size at up to 300 genome sample size (Fig. 4 and 5).

Quantities of genes in the core and pan-genome of phylotype D and B1 (N_D or N_{B1}) were expressed as a functions of number of analyzed genomes(g). Fitted equations for the core genome

curves of phylotype D (2.1 – according to Tettelin et al. 2005 model and 2.2 – according to Willenbrock et al. 2008 model):

$$N_D = 2001 + 1854\exp(\frac{-g}{42.3}) \quad (2.1)$$

$$N_D = 1090 + 6743\exp(\frac{-\sqrt{\sqrt{g}}}{4.38}) \quad (2.2)$$

Fitted equations for the core genome curves of phylotype B1 (2.3 – according to Tettelin et al. 2005 model and 2.4 – according to Willenbrock et al. 2008 model):

$$N_{B1} = 2733 + 1383\exp(\frac{-g}{20.96}) \quad (2.3)$$

$$N_{B1} = 80 + 5401\exp(\frac{-\sqrt{\sqrt{g}}}{3.99}) \quad (2.4)$$

Fitted equations for the pan-genome curve of phylotype D and B1, respectively:

$$N_D = 5013 + 16.8(g-1) + 178\exp(\frac{-2}{19.26})\frac{1-\exp(\frac{-(g-1)}{19.26})}{1-\exp(\frac{-1}{19.26})} \quad (2.5)$$

$$N_{B1} = 4620 + 30.4(g-1) + 192\exp(\frac{-2}{12.47})\frac{1-\exp(\frac{-(g-1)}{12.47})}{1-\exp(\frac{-1}{12.47})} \quad (2.6)$$

Table 4. Summary of gene content variation.

Phylotype	Size of pan genome	Size of core genome	Number of accessory genes per genome	Quantity of variants used for GWAS	Average genome
D	18,430	2,003	2,658.4	2,097	4,661
B1	13,374	2,761	1,804.6	1,420	4,565

Fitted equations predict that at 300 genomes the core and pan-genome of phylotype B1 are larger than phylotype D at 1.36 and 1.2 times, respectively (Fig. 4, 5). This observation seems to be counterintuitive considering that in phylotype D size of the average genome, size of the pan-genome (at 56 genome, Fig. 4, 5), and the average number of accessory genes per genome was

larger than in phylotype B1. This inconsistency in part can be explained by larger content of singletons (unique genes) in accessory part of genomes of phylotype B1 – 48.4%, compared to the range from 33% to 38% in D (based on three random subsamples of 56 genomes). In other words, the pan-genomes of phylotypes B1 and D consisted of 48% and 35% of singletons, respectively. Bigger amount of singletons in B1 possibly is a byproduct of higher recombination rate indicating that this phylotype is in more active phase of acquiring of new genes. This is consistent with idea that phylotype B1 is possibly in the early stage of divergence and speciation (Didelot et al. 2012). In a study of the pan-genome of *Prevotella* quantity of singleton genes reached 69% (Gupta et al. 2015). Gupta and colleagues did not find any evidences that unique genes are associated with ecological niche. Functionally, singletons in *Prevotella* belong to genome repair and replication, membrane synthesis and transcription classes. Still, most of singletons in *Prevotella* are genes of uncharacterized and hypothetical protein of unknown function. Reasons why such huge amount of the unique genes exists in some microbial genomes are yet to be studied (Zhang and Sievert 2014).

Previously it was shown on dataset of 20 genomes of *E. coli* that pan-genome of this species is "open" (Touchon et al. 2009). Similar estimation was obtained in another study on 27 genomes of *E. coli* (Didelot et al. 2012). Both studies were focused on commensal, pathogenic and laboratory strains of *E. coli*. According to our results on environmental isolates of *E. coli* the size of pan-genome of both phylotypes continues to grow at 300 genomes sample size. Consequently, it is possible that size of pan-genome in *E. coli* is unlimited. For this reason exchange of genetic information in bacteria *via* horizontal gene transfer is indeed very important in evolution of architecture of microbial genomes (Gogarten et al. 2002).

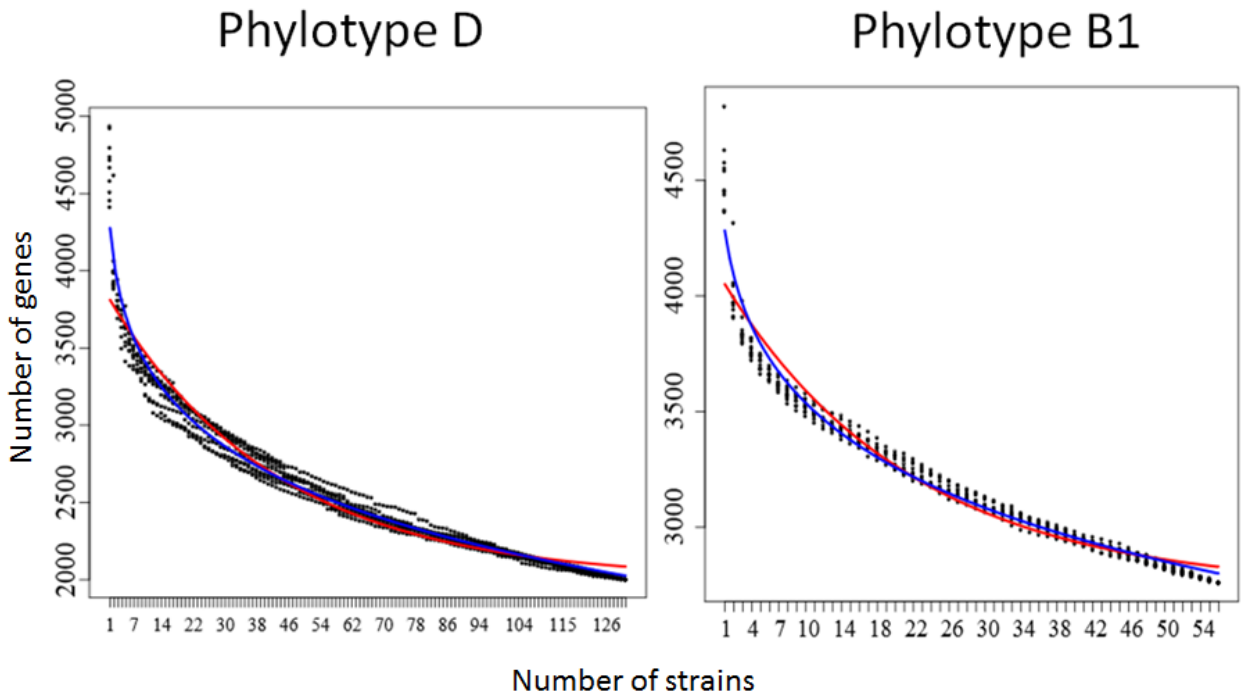

Figure 2. Estimation of the core genome size in phylotype D and B1. The number of specific genes (y-axis) is plotted as a function of the number of strains (x-axis) sequentially added. GET_HOMOLOGUES software performs 10 random input orders of avaliable set of genomes and then fits regression line. Equations for phylotype D: 2.1 (red line) and 2.2 (blue line). Equations for phylotype B1: 2.3 (red line) and 2.4 (blue line).

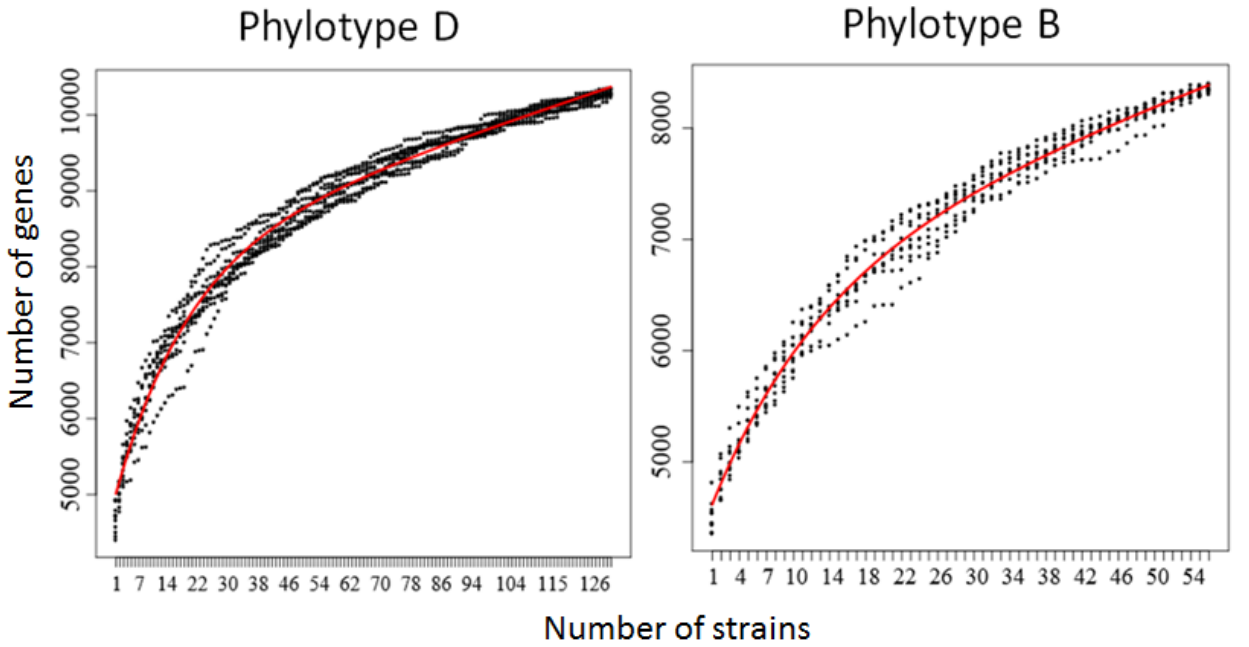

Figure 3. Estimation of pan-genome size in phylotype D and B1. The number of specific genes (y-axis) is plotted as a function of the number of strains (x-axis) sequentially added. GET_HOMOLOGUES software performs 10 random input orders of avaliable set of genomes and then fits regression line (equation for phylotype D – 2.5; equation for phylotype B1 – 2.6).

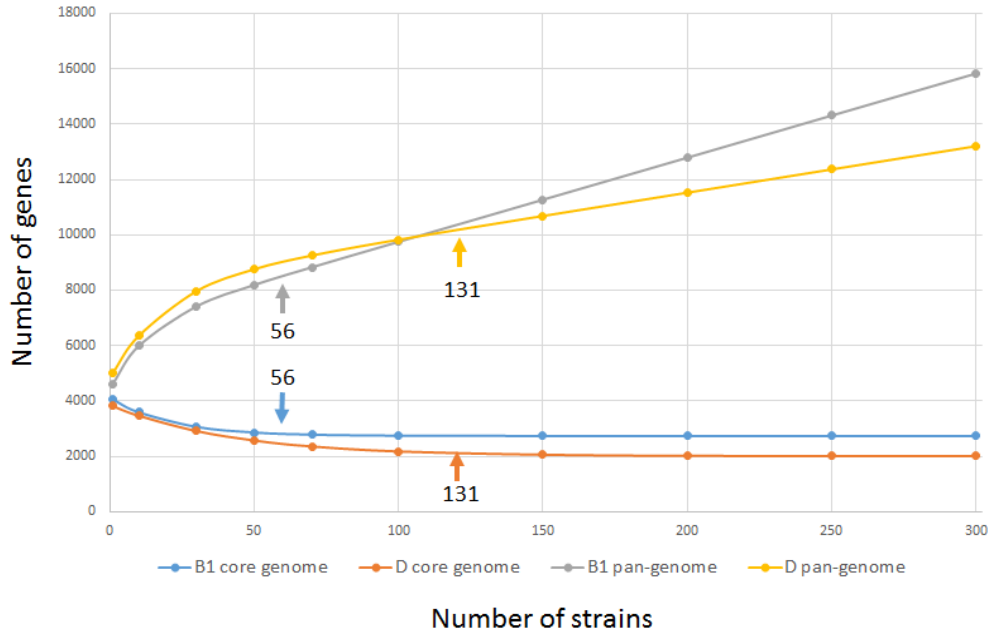

Figure 4. Extrapolation of the core and pan-genome sizes for phylotype D and B1 with increasing sample sizes of sequenced genomes. Arrows indicate actual quantity of isolates used in this study.

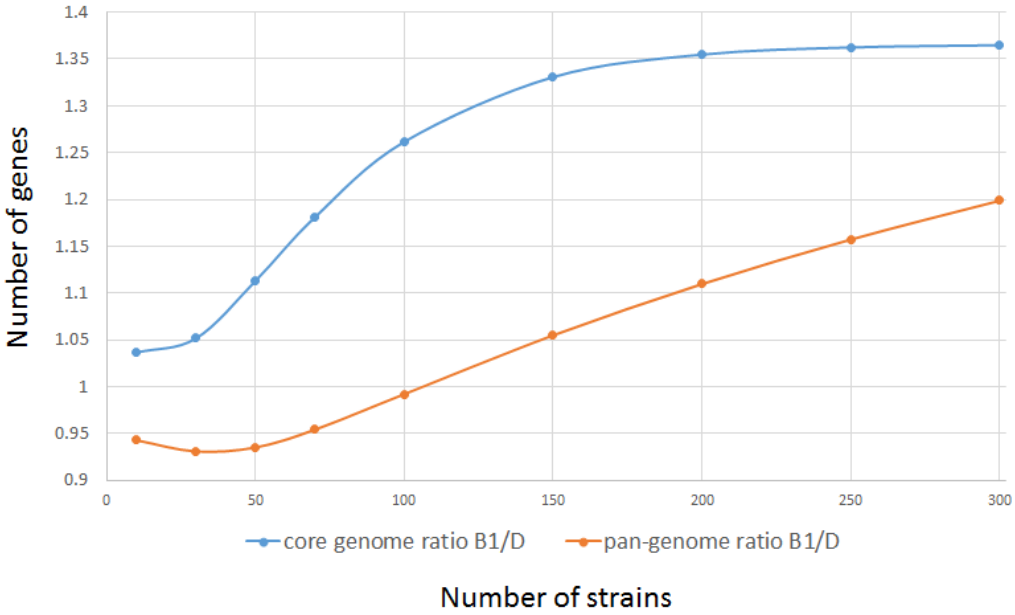

Figure 5. Relative difference of the core genome and pan-genome between phylotypes D and B1 with increasing sample sizes of sequenced genomes. Values for this plot were calculated from extrapolation of the core and pan genome sizes for phylotype D and B1 (Fig. 4) by dividing the number of genes in core genome or pan-genome of phylotype B1 by the number of genes in core genome or pan-genome of phylotype D.

Variation in the core genome

Using cortex_var software I identified 28,105 of SNPs and indels in the core genome of phylotype D; in phylotype B1 – 36,353. Using GATK software, I identified 364,114 SNPs/indels in phylotype D and 123,328 in B1. Inconsistency of the amount of variation identified by different callers might be explained by different sample size of phylotypes, different size of the core genome (Table 4). It was also repeatedly shown that SNP/indel calling is inconsistent between different software (Liu et al. 2013; Yi et al. 2014). Reason for inconsistency between variant callings originate from different stringency of filtering of mapping errors.

I identified 5,593 and 16,763 SNPs and indels in phylotype D and B1, respectively (Table 5) that were called by both SNP and indel callers: cortex_var and GATK. For random forest I used variants that were only present in more than 10% and less than 90% of genomes (Table 5). Quantity of SNPs/indels in phylotype B1 was higher than in phylotype D probably because phylotype B1 had larger core genome (Table 4). However, this is probably a partial explanation because the quantity of SNPs and indels in phylotype B1 was 3-times larger than in phylotype D, while size of the core genome was only 1.4 times larger. Still, an increase of sample size of B1 would probably reduce to some extent the size of the core genome and consequently reduce the number of SNPs/indels.

Table 5. Summary of variation in the core genome.

Phylotype	Number of strains	Total number of different SNP/indels across all genomes	Average quantity of SNP/indels per genome	Quantity of SNP/indels used for GWAS
D	131	5,593	706.2	1,232
B1	56	16,763	2,869.6	5,860

Among 5,593 variants in phylotype D I identified 893 intergenic variants, 3,697 synonymous variants, 959 missense variants, 29 inframe and frameshift indels; 10 nonsense; 4 stop-retained variants; 2 stop lost variants. In phylotype B1 I identified 1,745 intergenic variants; 10,925 synonymous variants; 3,939 missense variants; 77 inframe and frameshift indels; 55 nonsense variants; 12 stop-retained variants; 5 stop lost variants, 5 start lost variants.

To investigate the population structure of strains from different habitats I calculated an average protein identity (AAI) between each pair of strains within phylotype. The AAI across shared proteins varied in range from 98.7% up to 99.9% in phylotype D and from 99.4% up to 99.9% in phylotype B1. The AAI and average nucleotide identity are frequently used to identify species and investigate clustering (i.e. subpopulations) within species (Konstantinidis and Tiedje 2007). To visualize relatedness between studied isolates I generated heatmap plots based on Manhattan distance calculated from AAI (Fig. 6 and 7). I hypothesized that strains from same environment will have similar AAI either because of common evolutionary past or similar adaptive variation. Both phylotypes have clusters of strains that have higher AAI. However, strains did not cluster together (Fig. 6 and 7) by same habitat or by location from where they originate (second column with colors near dendrogram in Fig. 6 and third column with colors in Fig. 7). Subtypes of B1 phylotype: B1A, B1B, B1C also did not cluster together by the same metric (Fig. 7). Phylotype B1 had 4 subclusters and area of potentially high recombination (upper right corner of plot (Fig. 7). Phylotype B1 is already known for being implicated in frequent recombination/hybridization within B1 lineage and with phylotypes A and D (Didelot et al. 2012). Still, spatial isolation, different ecological niches (e. g. different strength of association with host or environmental habitat) and adaptive selection maintain existing divergence between different phylotypes of *E. coli* and force them to diverge further despite the recombination and hybridization

events between them (Didelot and Maiden 2010; Fraser et al. 2007; Gordon et al. 2002; Vulic et al. 1997).

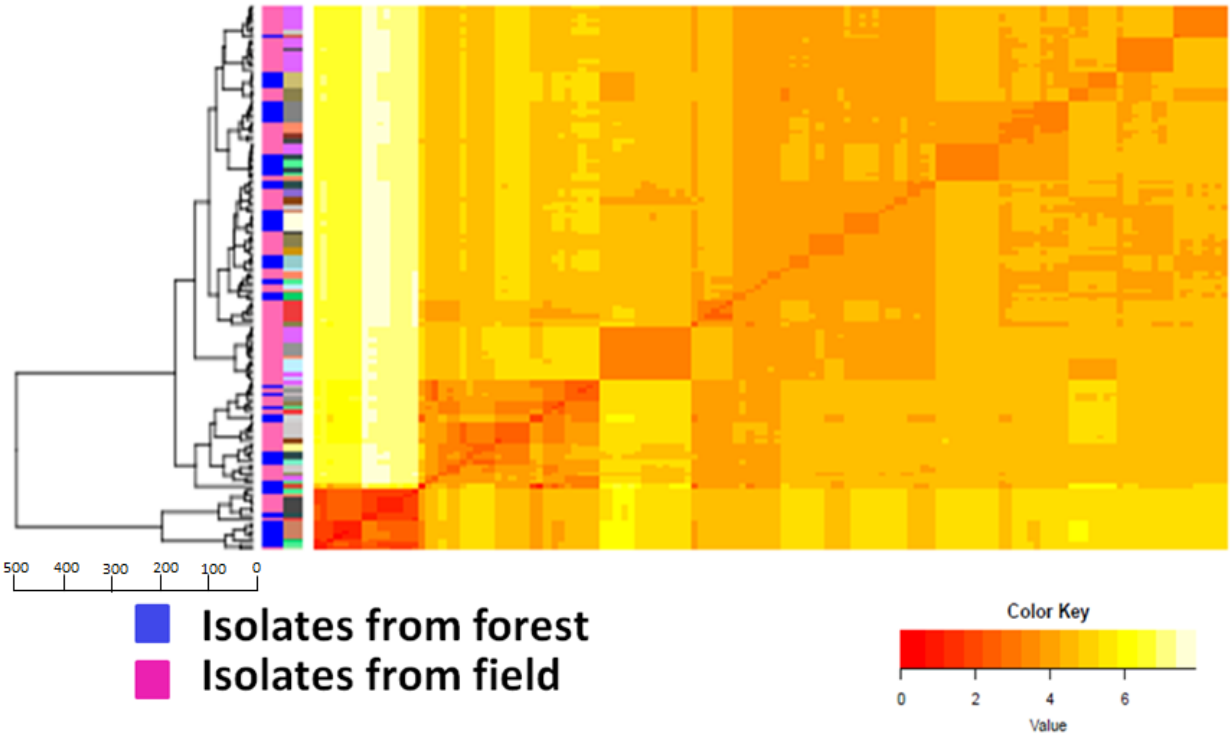

Figure 6. Heatmap of Manhattan distance calculated based on the average protein identity between pairs of strains of phylotype D. Second column of colors near dendrogram represents sampling sites from where strains originate.

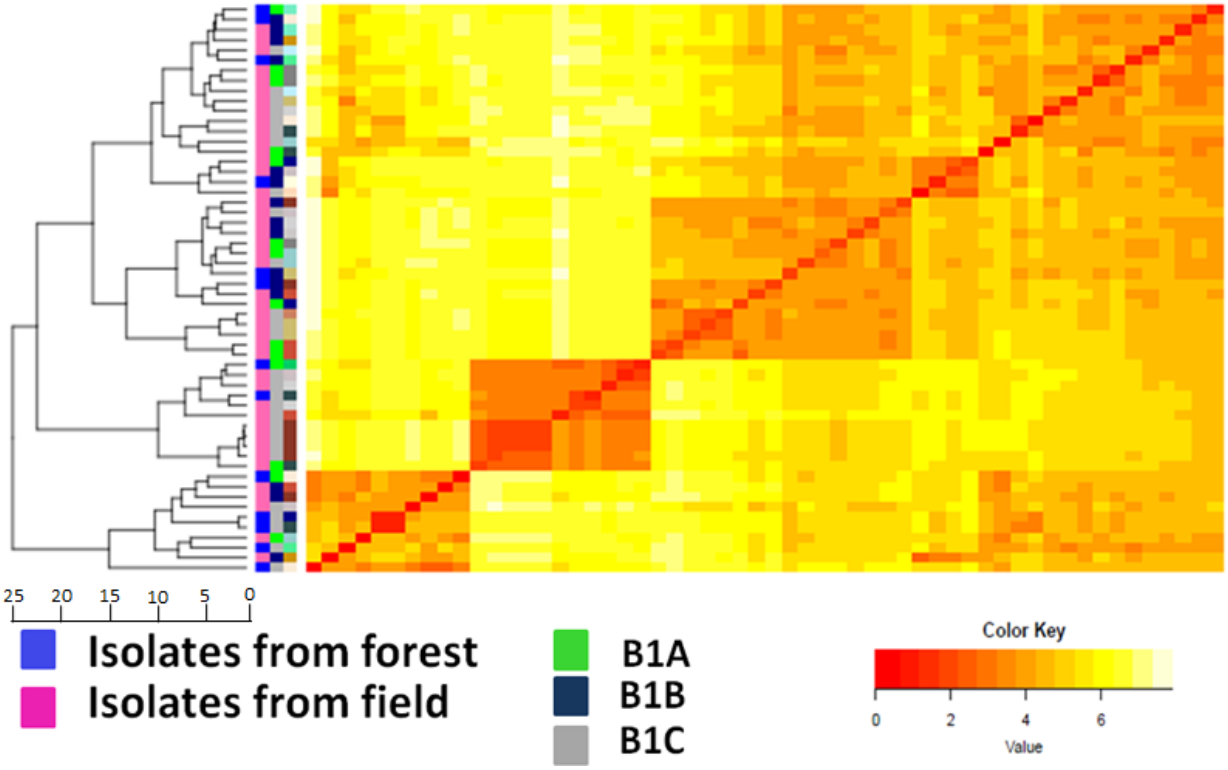

Figure 7. Heatmap of Manhattan distance calculated based on the average protein identity between pairs of strains of phylotype B1. The second column of colors near dendrogram represents sampling sites from where strains originate. The third column of colors near dendrogram represents sampling sites from where strains originate.

Evolutionary models of the ecotype divergence assume different strength of interaction between variation in the core genome and accessory genome (Cohan 2005). According to stable ecotype model core genome variants are major driving force of ecotype evolution. Recursive niche invasion model suggests that accessory genes are more important in the ecotype divergence (Godreuil et al. 2005). Nano-niche invasion model assumes that divergence into ecotypes is slow and subject to a complex interplay between accessory and core genomes (Cohan 2005). If stable ecotype or niche invasion models are true than we will observe very limited reassortment between the core and accessory variation. To investigate co-clustering of variation in the core and accessory parts of genome I built dendrograms based on binary data of presence of SNPs/indels and accessory genes using Canberra distance (weighted version of Manhattan distance). Alignment of dendrograms (Fig. 8, 9) shows how much reassortment occurred between core and accessory

genomes (i.e. how independently core and accessory variants are accumulated in strains). Measure of reassortment between two dendrograms is entanglement value. Entanglement varies in range from 0 to 1 (0 – no reassortment between dendrograms; 1 – two dendrograms are completely different). Alignment of dendrograms shows that level of reassortment in phylotype B1 is higher than in phylotype D (entanglement value is higher) (Fig. 8, 9). Considering smaller range of variation of the average amino acids identity of strains in phylotype D and lower reassortment level than in phylotype B1, ecotype evolution in D may resemble stable ecotype model. Phylotype B1 had bigger reassortment between core and accessory parts of genome. For this reason it is possible that nano-niche model is better candidate to describe divergence of ecotypes in phylotype B1.

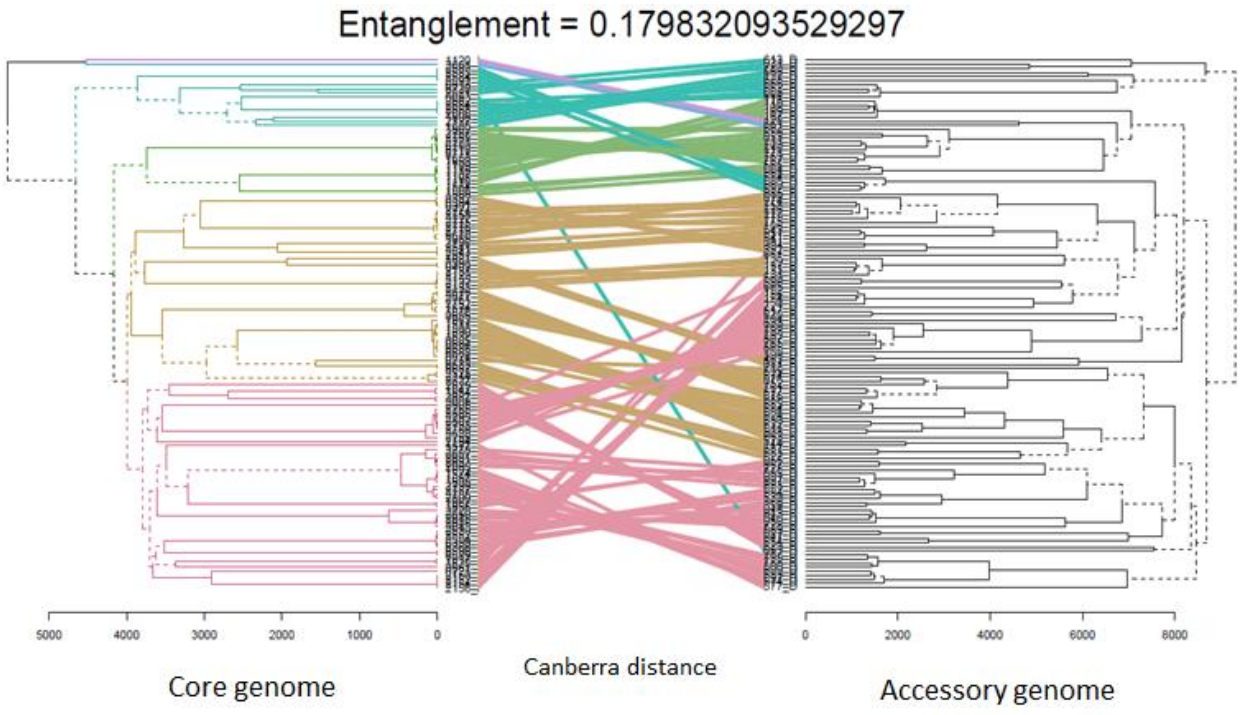

Figure 8. Tanglegram of variation in the core (left dendrogram) and accessory (right dendrogram) parts of genomes in phylotype D. Coloring represents groups of strains of the same cluster. Dashed lines indicate parts of dendrogram that were twisted to find better alignment between dendrograms.

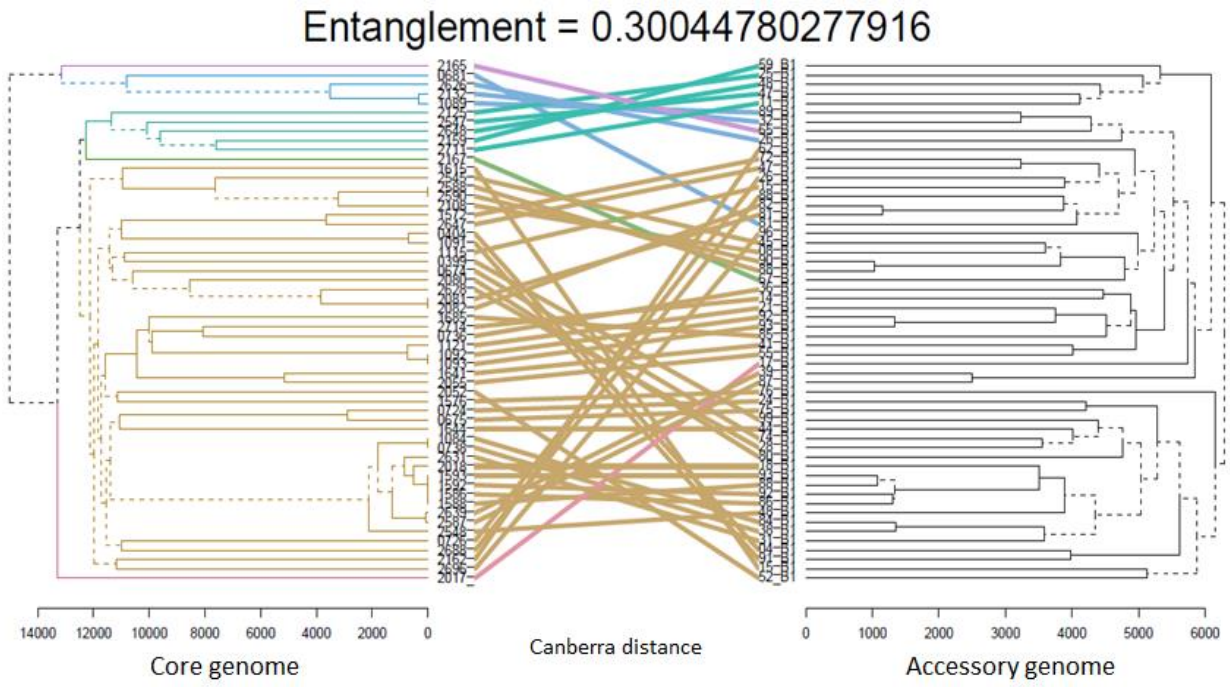

Figure 9. Tanglegram of variation in the core (left dendrogram) and accessory (right dendrogram) parts of genome in phylotype B1. Coloring represents groups of strains of the same cluster. Dashed lines indicate parts of dendrogram that were twisted to find better alignment between dendrograms.

Genome-wide search for variants associated with habitat

To find association between variation in *E. coli* genome with habitat I used random forest (RF) analysis. Performance of RF is presented in Tables 6 and 7 for phylotype D and in Tables 8 and 9 for B1. Balanced accuracy for class (field or forest) prediction based on the set of genomic variants in phylotype D was 84%. I also estimated RF performance when the same set of strains habitat was assigned randomly after permutation of actual set of habitat annotations. On permutated data balanced accuracy was 45%.

Table 6. Summary of RF performance on actual data from phylotype D.

Actual data	Habitat	Predicted class		Class error
		Field	Forest	
Actual class	Field	78	7	0.082
	Forest	11	35	0.239

Table 7. Summary of RF performance on randomized data from phylotype D.

Permutated data	Habitat	Predicted class		Class error
		Field	Forest	
Actual class	Field	64	21	0.2471
	Forest	39	7	0.8478

Due to small and unbalanced sample of forest compared to field isolates in phylotype B1 I were not able to obtain good random forest results. Balanced accuracy was 45% and 46% for actual and random data (also see Tables 8 and 9), respectively. Consequently variants identified by RF as important for habitat prediction in phylotype B1 are likely unreliable. I didn't consider results of RF for phylotype B1 for the downstream interpretation.

Table 8. Summary of RF performance on actual data from phylotype B1.

Actual data	Class	Predicted class		Class error
		Field	Forest	
Actual class	Field	39	4	0.09302326
	Forest	11	2	0.84615385

Table 9. Summary of RF performance on randomized data from phylotype B1.

Randomized data	Class	Predicted class		Class error
		Field	Forest	
Actual class	Field	40	3	0.069767
	Forest	13	0	1

In phylotype D top 5% (~170) of variants with the highest mean decrease of accuracy (MDA) were considered to be potentially associated with field or forest habitats. Cut-off was arbitrarily selected based on the previous research (Dutilh et al. 2014b) and Fig. 10. Range of MDA in top 170 variants declined from 0.0053675 (variant #1) to 0.0005355 (variant #170).

Relative presence of variants(P) in different habitats is present in Fig. 10 calculated using equation 2.7:

$$P = \frac{n_{field}}{N_{field\ total}} - \frac{n_{forest}}{N_{forest\ total}} \quad (2.7)$$

were n_{field} and n_{forest}, number of strains in which variant is present with respect to habitat; $N_{field\ total} = 85$ and $N_{forest\ total} = 46$. The core and accessory genome variants in top 5% of variants are evenly distributed between habitats (Fig. 10). Non of identified variants were absolutely habitat specific.

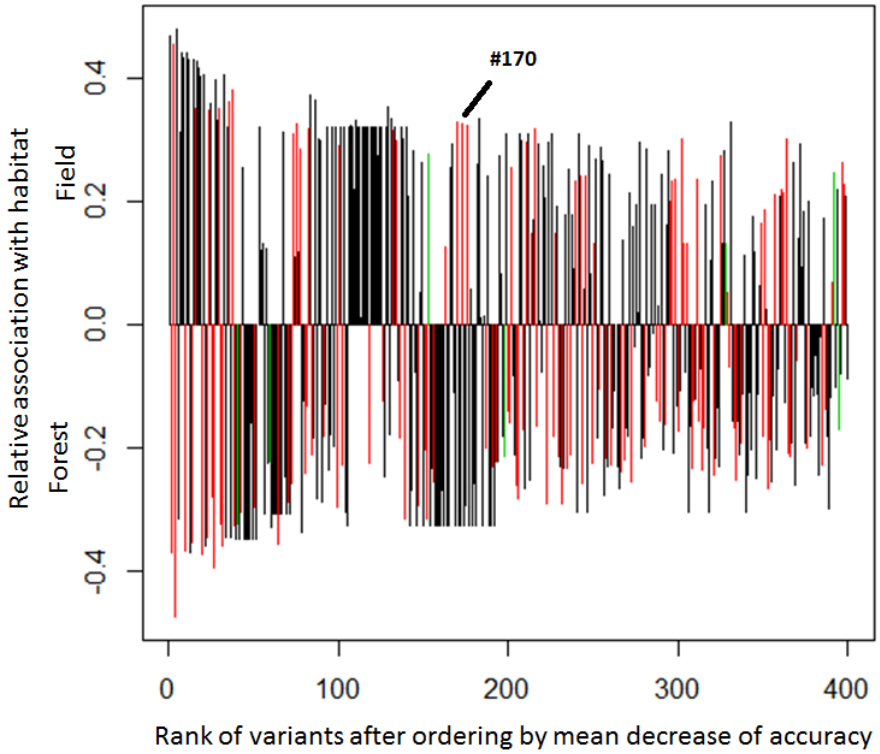

Figure 10. Relative presence of variants (y-axis) in isolates from the field (positive part of y-axis) or forest (negative part of y-axis). Variants are ordered by mean decrease of accuracy (x-axis represents rank after ordering). Black color – accessory gene variants; red – SNPs and indels; light green – intergenic variants.

In the top 170 important variants I identified 42 synonymous, 3 missense, 3 intergenic and 122 accessory gene variants. Presence of 66 accessory genes was associated with the field and 56 with the forest habitats.

Variation in accessory genome of field isolates

Isolates from the field contained 36 prophage genes and their derevatives (from several bacteriophages), 4 mobile element proteins, and 13 hypothetical proteins. Genes involved in cell wall biosyntheis (penicillin-binding protein, outer memrane lipoprotein Blc precursor) and nucleotide sugar/polyketide sugar biosynthesis (glucose-1-phosphate thymidylyltransferase and DTDP-glucose 4,6-dehydratase) were more prelevant in field isolates. HipA protein (from HipA-HipB toxin antitoxin (TA) system) was more frequent in field isolates. HipA toxin component of HipA-HipB type II TA system was serine/threonine-protein kinase that is involved in stringent response *via* RelA/SpoT and increases ppGpp levels, which inhibits cell wall synthesis, replication, transcription, and translation, reducing growth and leading to dormancy induction and drug resistance (Germain et al. 2013). The toxin-antitoxin systems are primarily known as selfish genetic systems which can cause postsegregational killing of daughter cells that have lost genome regions or plasmid containing these systems (Van Melderen and De Bast 2009). Toxin is a protein that has deleterious effect on cell functioning. Antitoxin that inhibits toxin can be either an antisense RNA to the toxin mRNA (type I and III TA systems) or a protein (type II, IV, V) (Ramage et al. 2009). It is suggested that TA systems are involved in persistence under starvation/stress conditions, growth control, and restriction of genome deterioration (Van Melderen 2010; Van Melderen and De Bast 2009).

Agricultural field soil possibly is less hospitable environment than forest soil because of the daily tempreature fluctuation, propencity to dessication, treatment with pesticides, and eutrophication (Cambardella et al. 1994; Crowther et al. 2014; Kuramae et al. 2012). It was previously observed that eutrophication in water ecosystems lead to mass overreproduction of microrganisms and consequent extinction which in part is ensured by transition of prophages to

lytic cycle (Kyle and Ferris 2013; Maurice et al. 2013; Ricciardi-Rigault et al. 2000). Eutrophication of agricultural soil can also cause these changes due to useage of fertilizers (Ochoa-Hueso et al. 2014). I speculate that bacteria in agricultural field have tendency to harbor prophages that have similar ecological function as prophages in the water ecosystems. Possibly for this reason, prophages' genes were more prelevant in *E. coli* isolates from field.

Variation in accessory genome of forest isolates

Flagellar synthesis accessory genes (18 flagella structural proteins and sigma factor that regulates flagelar proteins' operon), 17 hypothetical proteins, and 7 uncharachterized proteins were the most overrepresented classes of genes in forest isolates among important variants. Flagella is involved in biofilm formation, cell adhesion (bacteria-bacteria, bacteria-host), and capability to migrate a short distanse (Haiko and Westerlund-Wikstrom 2013). Possibly, some of these functions are particularly relevant for ecological niche of *E. coli* in forest soil. Also, forest isolates contained periplasmic phosphate ABC transporter, glycerol-3-phosphate cytidyltransferase, lysine-N-methylase, and MazEF type II TA system. MazEF TA system is involved in induction programmed cell death in bacteria (Engelberg-Kulka et al. 2005). Interestingly, a bacterial culture of *E. coli* experimentally deficient for MazEF TA-system produced more particles of phage P1 than parental wild type *E. coli* with functional MazEF TA-system (Hazan and Engelberg-Kulka 2004). I observed that prophage genes' were less abundant in the isolates from forest among 170 most important variants. It is possible that presenece of MazEF TA system in *E. coli* population reduces phage content, making them less successful in the microbial community forest soil.

Isolates from the forest soil contained elements from type VI secretory system (Hcp, IcmF related protein, ImpG/VasA protein). In *E. coli* this secretory system was previously described in enteroaggregative strains (Dudley et al. 2006). Type VI secretory system (T6SS) is involved in

bacteria-bacteria and bacteria-eukaryotic host interaction (Russell et al. 2014). In particular, this secretory system is involved in delivery of the toxic proteins to target cell *via* a "knife-like" mechanism. Functional roles of T6SS at ecosystem/microbial community level are invasion or defense against invasion of the communities, signaling, killing "cheaters"/non-cooperators/phage-infected bacteria, and biofilm remodeling (Bingle et al. 2008; Russell et al. 2014). It was suggested that TA systems and T6SS are functionally overlapping systems in the sense of the community ecology and physiology of bacteria (Russell et al. 2014). TA systems and T6SS contribute to defense against bacteriphages, formation of biofilms, regulation of metabolism, and stress respose. According to my study, soil isolates of *E. coli* of phylotype D harbor at least three modulators of growth bacteriophages, TA systems and T6SS, that differ from field isolates. Importantly, prophage genes seem to be more associated with agricultural field habitat, while TA systems and T6SS are associtaed with forest habitat. This biased distribution of the accessory variants across habitats may be caused by 1) antagonistic interactions between TA systems and T6SS on one side and bacteriophages on the other side and/or 2) beneficial effects of prophages in *E. coli* from agricultural field and beneficial effects of TA systems and T6SS in isolates from forest soil.

Variation in the core genome

Distribution of all important variants in the core genome are shown in Fig. 11. Only one missense substitution in the gene *pabB* (aminodeoxychorismate synthase, subunit I) was associated with the field habitat and two missense substitutions in the genes *yliC* (putative oligopeptide transporter) and *ypdF* (aminopeptidase) were associated with the forest habitat. Variation in the peptide utilization and amino acid utilization genes were previously shown to be associated with survival in non-host environments in *E. coli* and *Salmonella* (Winfield and Groisman 2003).

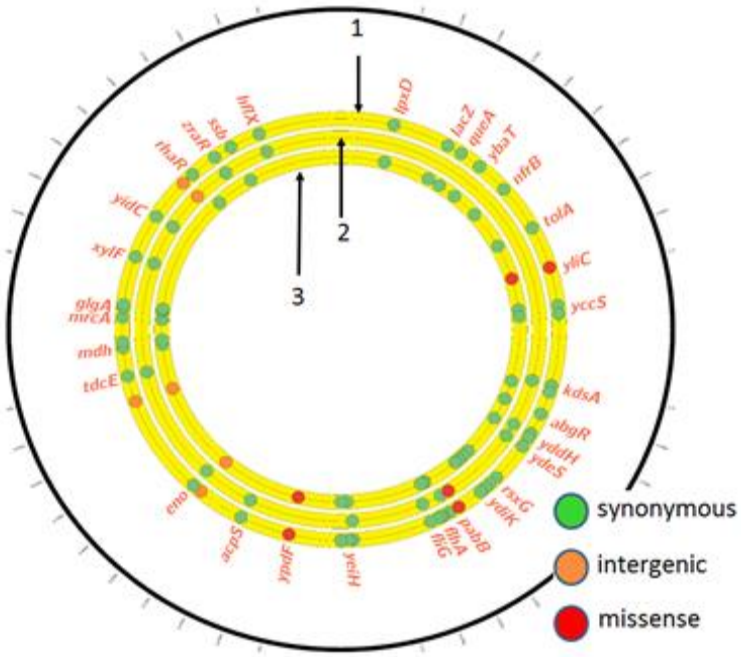

Figure 11. Plot of variants that were more associated with soil from field or forest. 1 – all important variants, 2 – variants associated with field, 3 – variants associated with forest.

Most of the observed variants that were associated with habitat are synonymous (Fig. 11). Genes with synonymous substitutions belong to stress response; energy metabolism; amino acids and sugars metabolism, ion transport, etc (Table 10, 11). Previous studies observed similar patterns of polymorphisms in isolates from extrahost environment (van Elsas et al. 2011a). Synonymous variation possibly is a reflection of accumulation of neutral variation rather than adaptive variation (Maddamsetti et al. 2015). Still, synonymous variation can contribute to adaptation *via* changing functionality of regulatory element structure and/or gene expression, leaving the primary protein structure intact (Plotkin and Kudla 2011).

Table 10. Single nucletide polymorphisms discovered in phylotype D associated with forest.

Gene name	Gene ID	Mean decrease of accuracy	Function	GO process
yccS	ECUMN_1150	0.005283101	predicted inner membrane protein	NA
mdtK	ECUMN_1953	0.004494163	multidrug efflux transporter MdtK	dipeptide transmembrane transport
abgR	ECUMN_1635	0.002881105	predicted DNA-binding transcriptional regulator, LYSR-type	regulation of transcription, DNA-templated
tap	ECUMN_2182	0.002535147	TAP-MONOMER	signal transduction
glpD	ECUMN_3884	0.002182674	GlpD	anaerobic respiration
glgA	ECUMN_3893	0.001647375	glycogen synthase	cellular response to DNA damage stimulus
mrcA	ECUMN_3854	0.001573429	EG10748-MONOMER	cell wall organization
fdnG	ECUMN_1727	0.001571372	formate dehydrogenase N, α subunit	cellular respiration
flhA	ECUMN_2176	0.001399613	flagellar biosynthesis protein FlhA	bacterial-type flagellum organization
btuD	ECUMN_2000	0.001394905	vitamin B₁₂ ABC transporter – ATP binding subunit	vitamin transmembrane transport
yeiR	ECUMN_2510	0.001225219	zinc-binding GTPase	NA
napB	ECUMN_2538	0.001175152	subunit of periplasmic nitrate reductase, cytochrome c₅₅₀ protein	oxidation-reduction process
mdh	ECUMN_3710	0.00110592	malate dehydrogenase	carboxylic acid metabolic process
accC	ECUMN_3730	0.000942635	AccC	fatty acid metabolic process
agp	ECUMN_1184	0.000910501	glucose-1-phosphatase	glucose catabolic process
ybaT	ECUMN_0525	0.000900944	putative transport protein, archaeal/bacterial transporter (ABT) family	L-alpha-amino acid transmembrane transport
adhE	ECUMN_1538	0.000818069	AdhE	metabolic process

Table 10. Single nucletide polymorphisms discovered in phylotype D associated with forest (continued).

Gene name	Gene ID	Mean decrease of accuracy	Function	GO process
ydiK	ECUMN_1977	0.000808288	predicted inner membrane protein	NA
nfrB	ECUMN_0642	0.00077204	bacteriophage N4 receptor, inner membrane subunit	transport
	ECUMN_4435	0.000726951	RhaR transcriptional activator	positive regulation of transcription, DNA-templated
araG	ECUMN_2194	0.000683487	arabinose ABC transporter - ATP binding subunit	metabolic process
queA	ECUMN_0443	0.000680751	EG10812-MONOMER	tRNA wobble guanine modification
rsxG	ECUMN_1922	0.000639093	member of SoxR-reducing complex	oxidation-reduction process
glgA	ECUMN_3893	0.000618777	glycogen synthase	cellular response to DNA damage stimulus
ypdF	ECUMN_2715	0.000603961	aminopeptidase	proteolysis
ssb	ECUMN_4595	0.000584278	SSB monomer	positive regulation of catalytic activity
lacZ	ECUMN_0387	0.000573996	β-galactosidase monomer	metabolic process
yliC	ECUMN_1019	0.000566535	Putative peptide transporter permease subunit: membrane component of ABC superfamily	NA
tolA	ECUMN_0827	0.000566307	TolA – inner membrane protein of the Tol-Pal system	protein import
lpxD	ECUMN_0176	0.000560825	UDP-3-O-(3-hydroxymyristoyl)glucosamine acyltransferase	lipid metabolic process

Table 11. Single nucletide polymorphisms discovered in phylotype D associated with field.

Gene name	Gene ID	Mean decrease of accuracy	Function	GO process
fliG	ECUMN_2231	0.00471207	flagellar motor switch protein FliG	metabolic process, bacterial-type flagellum-dependent cell motility
tdcE	ECUMN_3598	0.00245461	2-ketobutyrate formate-lyase, pyruvate formate-lyase 4, 2-ketobutyrate formate-lyase/pyruvate formate-lyase 4, inactive	threonine catabolic process
yddH	ECUMN_1713	0.00162987	conserved protein	oxidation-reduction process
xylF	ECUMN_4077	0.00142867	xylose ABC transporter – periplasmic binding protein	carbohydrate transport
yidC	ECUMN_4237	0.00132843	inner-membrane protein insertion factor	protein homooligomerization , protein transport
eda	ECUMN_2147	0.00127291	Eda	metabolic process
acpS	ECUMN_2884	0.00088887	AcpS	fatty acid metabolic process
kdsA	ECUMN_1512	0.00087996	3-deoxy-D-manno-octulosonate 8-phosphate synthase	protein homotetramerization, lipopolysaccharide biosynthetic process
hflX	ECUMN_4706	0.00085581	GTPase associated with the 50S subunit of the ribosome	metabolic process, response to heat
eno	ECUMN_3110	0.0007836	Eno	glycolytic process
ydeS	ECUMN_1758	0.00068343	predicted fimbrial-like protein	cell adhesion
yeiH	ECUMN_2494	0.00060541	conserved inner membrane protein	NA
pabB	ECUMN_2104	0.00060485	Aminodeoxychorismate synthase, subunit I	folic acid-containing compound biosynthetic process
zraR	ECUMN_4528	0.00054776	ZraR-Phosphorylated DNA-binding transcriptional activator, ZraR transcriptional activator	regulation of transcription, DNA-templated, phosphorylation signal transduction system
nagD	ECUMN_0760	0.00053437	NA	NA

Various studies suggest that the accessory genes and horizontal gene transfer play key role in divergence of the bacteria into ecological niches (Dutilh et al. 2013; Dutilh et al. 2014b; Heuer and Smalla 2012; Niehus et al. 2015; Wiedenbeck and Cohan 2011) Typically, accessory genome is composed from prophages, mobile elements, flarellar transcription factors, antibiotic resistance genes (Pitout 2012), pathogenicity genes (Dobrindt et al. 2004), accessory metabolic genes (e. g. genes of terpenoids metabolism in *Pseudomonas auragenosa* (Kung et al. 2010; Mathee et al. 2008)). In our study the most abundant group of genes important to predict origin of isolates were prophage genes, mobile elements and hypothetical proteins genes. I observed small quantity of predominantly synonymous variation in the core genome associated with habitat. The accessory genes associated with habitat composed up to 72% of list of top 5% of the most important variants identified by RF. Considering smaller divergence based on the average amino acids identity of strains in phylotype D and lower reassortment level than in phylotype B1 ecotype evolution in D may resemble stable ecotype model. Phylotype B1 had bigger reassortment between core and accessory parts of genome for this reason it is possible that Nano-niche model is better candidate to describe divergence of ecotypes in phylotype B1.

Conclusions

1. Fitted regression lines obtained from estimation of the core and pan-genome size by sampling predict that the core and pan-genome are larger in phylotype B1 than in D. Pan-genomes of both phylotypes are possibly "open".
2. Phylotype B1 contains more singletons in the accessory genome than phylotype D.
3. Isolates don't cluster by habitat or sampling site based on the average amino acids identity.
4. In phylotype D, prophage accessory genes are more associated with field habitat while TA system and T6SS genes are more associated with forest habitat.

5. Smaller divergence based on the average amino acids identity of strains in phylotype D and lower reassortment level than in phylotype B1 suggests that ecotype evolution in D may resemble Stable ecotype model. Phylotype B1 had bigger reassortment between the core and accessory parts of genome for this reason it is possible that nano-niche model is better candidate to describe divergence of ecotypes in phylotype B1.

CHAPTER 3. VARIATION IN CORE AND ACCESSORY GENOME ASSOCIATED WITH SURVIVAL OF *E. COLI* IN SOIL

Introduction

New ecotypes arise frequently in microbial populations as a result of the biological combination of: a) enormous population sizes, b) fast mutation rates, c) acquisition of niche-modifying genes *via* horizontal gene transfer (HGT), d) plastic genomes which are prone to rearrangement and changes in gene regulatory networks, and e) passive dispersal which leads to continuous subdivision of microbial metapopulations into spatially-isolated subpopulations undergoing diverse regimes of environmental selection. Homologous recombination then provides a means for the spread of adaptive alleles to other lineages given that those subpopulations are connected by migration back into the metapopulation gene pool (Cohan 2005; Cordero and Polz 2014a). It was recently shown that intensive recombination within linage of species does not interfere with ecotype divergence. Moreover, reduction of recombination between ecotypes follows ecological divergence rather than precedes it (Melendrez et al. 2016).

Escherichia coli is no exception to this and provides a striking example of what can be learned (Brennan et al. 2010; Luo et al. 2011). *E. coli* is primarily a host-associated enteric species of bacteria. The species is represented primarily by commensal variants belonging to one of eight recognized phylotypes: A, B1, B2, C, D, E, F, and *Escherichia* cryptic clades I-IV (Clermont et al. 2013). In extrahost environment, phylotypes B1 and D are more prevalent than any other (Bergholz et al. 2011; Orsi et al. 2007; Ratajczak et al. 2010). Whereas phylotype A is found quite rarely in soil habitats, it is found more frequently in water. *E. coli* is frequently found to proliferate and/or persist on plants (Brandl 2006). It was shown that phylotypes B1 and D had strong

association phenotype with plants, A and E also showed association phenotype with plants but to lesser extent (Meric et al. 2013).

The causes of these associations with different extrahost habitats are not always clear, but the existence of association suggests that phylotype-level, and perhaps smaller ecological subdivisions exist within the *E. coli* species. Indeed, some groups may be adapted for growth outside the host (van Elsas et al. 2011b; Walk et al. 2009b) and are thought to confound the use of *E. coli* as a water quality indicator. So, this enormous biodiversity within a single microbial species is not just a challenge to evolutionary biologists. Rather, it also has policy implications for environmental quality monitoring and management of watersheds (Santo Domingo et al. 2007; Verhougstraete et al. 2015).

When microbial populations are subdivided by different environmental selection pressures, genome-wide association studies and population genomic analyses can be used to understand the generation, maintenance and spread of adaptations for extrahost survival, growth and/or dispersal (Cordero and Polz 2014b; Dutilh et al. 2014a; Shapiro et al. 2012).

Aims of this this study are to elucidate the potential for adaptive traits be selected by extrahost environments and to spread of fecal strains or *E. coli* pathogens with implications for both water quality and food safety. I hypothesized that frequent, rapid, and passive dispersal to soil is a key factor that selects genomic biodiversity in *E. coli*. To gain new insight into the role of dispersal into extrahost habitats in the selection and maintenance of diverse ecotypes, I conducted a study on select *E. coli* isolates from a recreational grassland soil in central New York State. I sought to address the questions: a) do *E. coli* phylotypes exhibit different capacities for persistence in soil, b) what adaptive variants are associated with enhanced persistence of *E. coli* in surface soils, and c) how diverse are adaptive variants in the core and accessory genomes? To answer these

questions, I performed genome-wide association study of variation in survival variation of 9 strains of each phylotype B1 and D (maximum death rate) that was measured in soil microcosms Hudson silt loam (pH 6.1, organic matter 7%) media.

Materials and Methods

Isolates origin and survival assay

Soil collection, isolation of *E. coli*, estimation of death rates and DNA-extraction were performed by research team at Food Safety Laboratory at Cornell University, NY: Laura Strawn, Steven Warchoki, Gina Ryan, Courtenay Simmons and Jihun Kang

Soil for *E. coli* isolation was collected at Mitchell St. Natural Area of the Cornell Plantations. Soil microcosms containing Hudson silt-loam were inoculated with *E. coli* strains at initial density of 10^2 CFU g^{-1} soil. Survival in microcosms was assayed by viable cell plating of 0.5 g soil following serial dilution in 1X PBS. At each sample day, 100 µl of three soil dilutions were spread plated onto EC medium with MUG agar plates and these plates were incubated at 37°C for 16-18 hours. After incubation, plates were observed under ultraviolet light to enumerate blue fluorescent colonies. Data were collected over 40 days with viable plate counts conducted on days 0, 1, 3, 5, 10, 20, 30, and 40. Log-transformed CFU g^{-1} measurements were fit to a four parameter Geeraerd logistic survival model (Geeraerd et al. 2000). The model estimates the lag-time (i.e., shoulder), maximum death rate (kmax), initial log10 CFU g^{-1} soil and minimum log10 CFU g^{-1}. Models were implemented using formulas from the nlstools package (Baty et al. 2015) in R 2.10.1 (R Core Team 2015). Two groups of *post hoc* survival phenotypes (group 1 – low and group 2 – high death rate) were assigned based on maximum death rate (kmax).

Genome resequencing and annotation

The genomes of phylotype D and clade B1B strains from the soil survival assay were sequenced to further explore the genomic variation associated with soil survival phenotypes.

Sequence libraries were generated using a Nextera Sequence Library preparation kit (Illumina, San Diego, CA) with oligonucleotide barcodes delineating each strain. Sequence reads of 100 bp size were obtained using an Illumina HiSeq 2500 at the Cornell Core Life Sciences Laboratories (http://www.biotech.cornell.edu/brc/genomics-facility). I performed quality control of paired-end reads using FastQC v. 0.11.2 (Andrews S. 2010). Trimmomatic v.0.32 (Bolger et al. 2014) was used to remove Nextera transposase sequence; remnants of adapters at the beginning of the reads (the first 10 nucleotides were removed in all reads); and bases with quality below threshold (Q=20) at the end of the reads. MAXINFO parameter was at 0.2 in favor to retain longer reads. All reads of length less than 60 bases were discarded. I applied Genome Analysis ToolKit v3.3 (McKenna et al. 2010) for SNPs and indels discovery and hard-filtering procedure taking into account best practices recommendations (DePristo et al. 2011). Reads for these analyses were mapped using stampy v.1.0.21 (Lunter and Goodson 2011) and bwa-mem v.0.6.2 (Li and Durbin 2009) on reference *E. coli* genomes of strains UMN026 and IAI1 for phylotypes D and B, respectively. Genomic variants were also determined by variant calling procedures using cortex_var v.1.0.5.21 (Iqbal et al. 2012). Annotation of SNPs and indels was performed using snpEff v.4.1 (Cingolani et al. 2012). Only variants with representation in all sequenced genomes and called by both GATK and cortex_var softwares were retained for downstream analysis. Genomic gene, phage and mobile element content were also examined *via de novo* assembly of the genome sequences using velvet v.1.1 (Zerbino and Birney 2008). *De novo* genome assembly statistics is in Table 13. *De novo* assemblies were annotated using Rapid Annotation using Subsystems Technology (RAST) service v.2.0 (Aziz et al. 2008). Gene content variation among strains was analyzed by BLAST search scripts (http://enveomics.blogspot.com). Briefly, amino acid coding sequences were queried against other genomes from the same subpopulation. A gene was scored as shared between two

genomes if BLAST search identified a homolog in the target genome that exhibited >= 80% amino acid sequence identity over >= 85% of the query sequence length.

Random forest algorithm was used to find variation in core and accessory parts of genome associated with *post-hoc* low and high death rates in survival experiment. Analysis was performed using randomForest package (Liaw and Wiener 2002) in R 3.2.0 3 (R Core Team 2015).

Results and Discussion

Improved survival in soil is likely due to both phenotypic plasticity (i.e., gene expression responses) and adaptive variation for soil survival with or without trade-offs (i.e., antagonistic pleiotropy) for fitness in host animals. To explore the basis for increased survival phenotypes in soil, I searched a small number of isolates for variants that could be associated with slow or fast death rates. Survival data were used to select nine strains from each of phylotypes B1B and D for sequencing. Using data on survival in soil microcosms, survival phenotypes were classified as low-death rate (group 1) and high death rate (group 2) within each phylotype. *Post hoc* assigned phenotype groups for Phylotype D had average death rates of 0.28±0.01 in group 1(low death rate) (n=5) and 0.33±0.01 in group 2 (high death rate) (n=4). For phylotype B1B, 0.21±0.03 in group 1 (low death rate) (n=5) and 0.36±0.15 in group 2 (high death rate) (n=4) (Fig. 12 and Table 12).

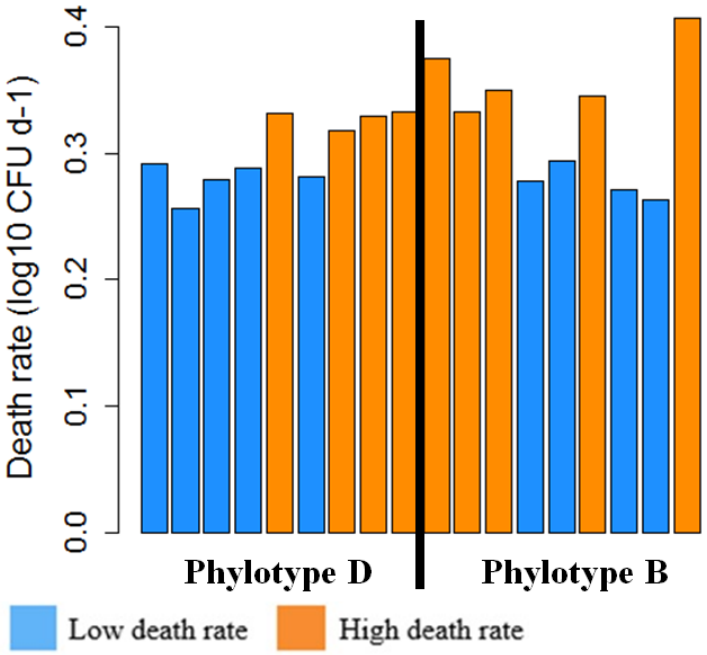

Figure 12. Death rates of *E. coli* strains in *post-hoc* high and low death rate classes by phylotype in Hudson silt loam soil microcosm (pH 6.0).

Table 12. Maximum death rate and post-hoc classification of 18 strains of *Escherichia coli*.

Phylotype	Isolate	Maximum death rate	Post-hoc classification
D	M1G6	0.29118	Low
D	J1D7	0.25639	Low
D	O2C11	0.27902	Low
D	O2F2	0.28760	Low
D	O2B1	0.33093	High
D	O4G6	0.28154	Low
D	J2H3	0.31791	High
D	J2F9	0.32929	High
D	J2E3	0.33218	High
B	J1A3	0.37425	High
B	J1E7	0.33217	High
B	J3A3	0.34978	High
B	J3D9	0.27834	Low
B	J3E2	0.29371	Low
B	O4E6	0.34532	High
B	O4F12	0.27088	Low
B	O4G2	0.26254	Low
B	O4G9	0.40623	High

Using *post hoc* phenotype designations, I sequenced genomes from each survival phenotypes group and from phylotypes D and B1B to discover associations between specific genomic variants and survival in surface soil. Genomes were sequenced to an average per-base coverage of 79-fold (range: 48.9-117.9 per-base coverage) (Table 13).

Table 13. *De novo* assembly statistics summary.

Isolate name	Total number of reads	Number of contigs	Median depth of coverage	N50-value	The longest contig size	Size of assembly	Number of Genes identified using RAST service
J1D7	7266686	121	48.9	322955	541674	4925483	4701
J2E3	13074084	128	81.3	207646	452103	4879163	4657
J2F9	14469716	188	96.2	180384	869913	5109392	4914
J2H3	12713888	147	83.1	126920	446591	4903242	4716
M1G6	12849364	326	58.4	132831	484830	5549144	5621
O2B1	10604384	109	76.7	305663	650334	5133645	4916
O2C11	14892698	96	86.7	187385	527157	4974171	4819
O2F2	10458192	216	63.6	193447	343344	5121460	4952
O4G6	11243738	119	70.2	291486	905073	4866139	4652
J1A3	13711202	100	100.7	192557	441084	4652138	4434
J1E7	9273348	173	63.9	159410	526399	4851427	4723
J3A3	11712888	77	83.8	182895	329543	4723822	4553
J3D9	14785500	101	117.9	215089	485377	4744688	4540
J3E2	11560682	99	86.95	195565	703389	4587420	4412
O4E4	9356984	161	56.8	157379	425934	4854679	4739
O4E6	11074730	151	67.2	184981	526412	4860526	4728
O4F12	8975222	112	72.2	162506	684788	4670372	4503
O4G2	13969122	106	110.9	159790	695395	4673355	4529
O4G9	10110714	109	67.8	162673	329465	4775586	4613

Average SNP content per genome was 4,977 with range from 433 up to 6,031 SNP. In total, 17,061 and 15,048 unique SNPs and indels were discovered in the core genomes of phylotypes D and B1B, respectively. Synonymous and intergenic variants dominated the SNP composition at 83% and 79% of SNP variants in D and B1B phylotypes, respectively. Missense variants composed 16% and 19% of SNP variants for D and B1B , respectively.

Survival in soil may be due to polymorphism in the core genome, gene content variation in the accessory (flexible) genome, or interactions between the two. In the two phylotypes, the total accessory genome consisted of 1,999 genes and 807 genes, respectively (Fig. 13).

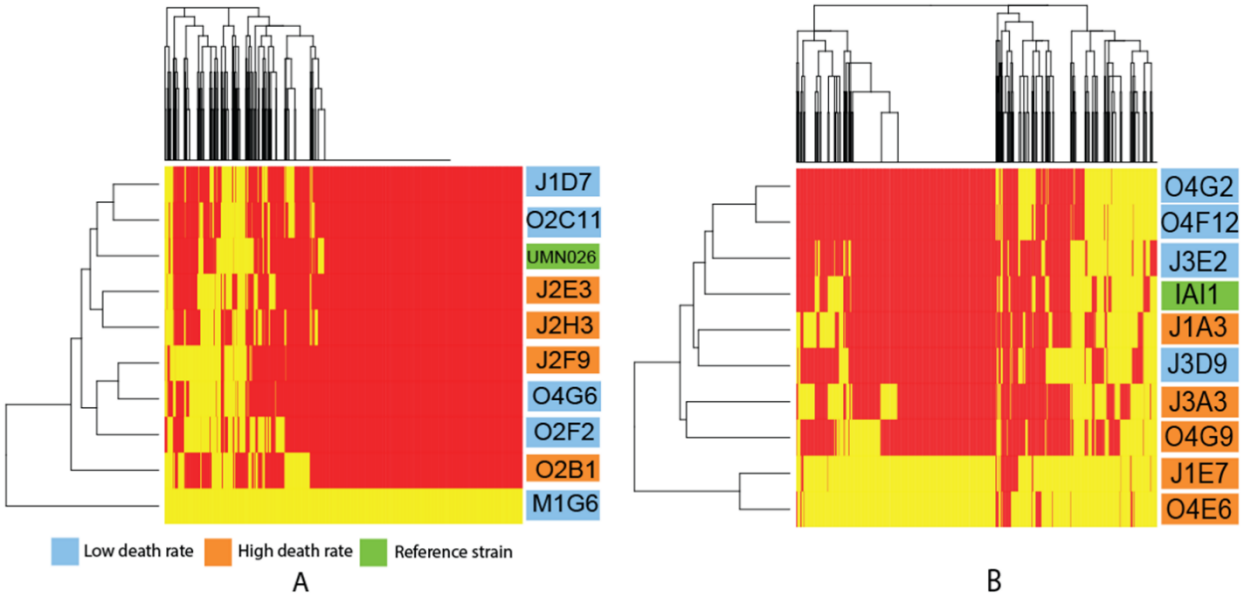

Figure 13. Heatmap of accessory genome in phylotype D (A) and B1B (B), yellow – gene is present, red – gene is absent. Strain names are colored according to survival group (blue - low death rate; orange - high death rate).

These accessory genome sizes are smaller than expected based on other *E. coli* data sets of similar size (Hendrickson 2009; Touchon et al. 2009), and I attributed the small accessory genome size to the close geographic and limited phylogenetic origins of the sequenced isolates. To examine co-clustering of core and accessory variants, I constructed dendrograms based on binary (presence/absence) matrices of core and accessory variants within phylotypes. Tanglegrams comparing the clustering of core and accessory variants showed that phylotype D strains exhibited little concordance between dendrograms, but phylotype B1B genome variants exhibited greater concordance between the two sets of variants (Fig. 14). There was limited reassorting of accessory and core genome variation in both phylotypes (Fig. 14).

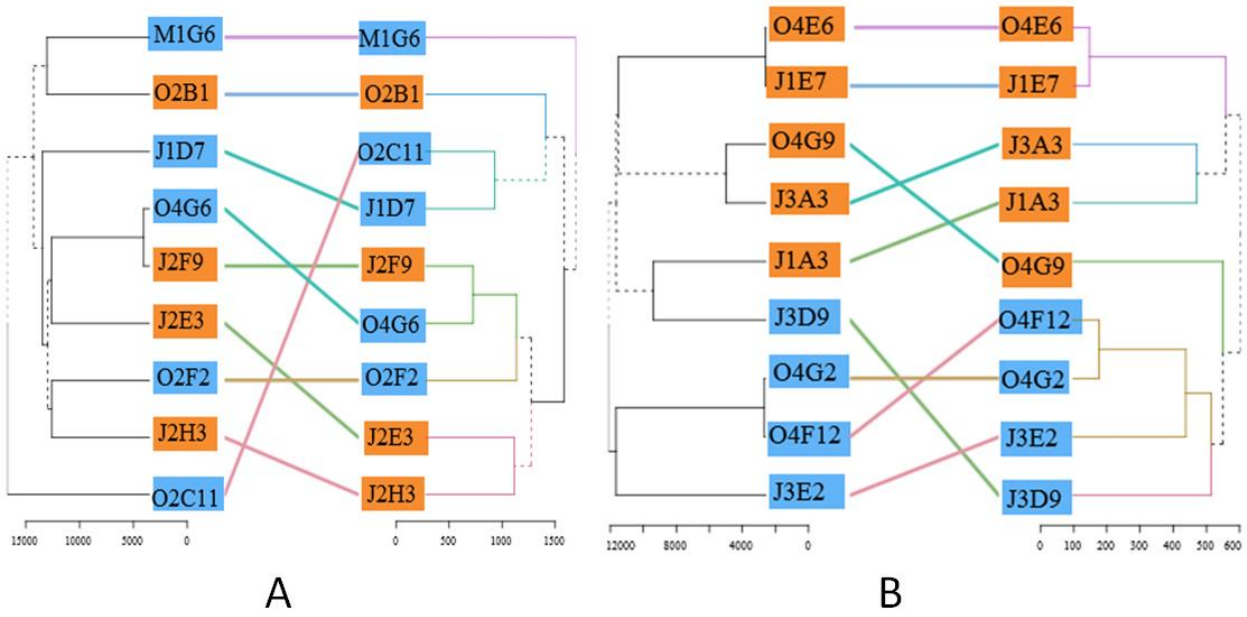

Figure 14. Tangelgrams of core (SNPs and indels) (left dendrogram) and accessory genomes variation (right dendrogram) in phylotype D (A) and B1B (B). Entanglement values: (A) E=0.23, (B) E=0.14 (orange – high death rate, blue – low death rate).

Random forests (RF) analysis was applied to link variants with maximum death rate (k_{max}) phenotype of sequenced strains (Dutilh et al. 2013; Dutilh et al. 2014b). Since the power of this analysis was very limited, due to small sample sizes, I focused on the top 100 most important variants identified *via* RF using the mean decrease in phenotype prediction accuracy criterion.

In phylotype D, missense variants among core genes *zraS*, *ltaE*, *frmB*, *abgB*, ECUMN_0533 (putative adhesin/invasin-like protein) had tendency to associate with high death rate phenotype (Table 14). However, association between genome variants and phenotype was less than clear, compared to phylotype B1B because one strain in phylotype D with high death rate contained each variant (Table 14). Random forest identified ten accessory genes that were associated with better survival. Among accessory genes in phylotype D strains, presence of prevent host death protein (Phd antitoxin) gene which is a part of Phd-Doc toxin-antitoxin and two hypothetical proteins were associated with low death rate.

Table 14. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype D.

Rank in top 100 RF hits	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of SNP by survival class		Gene/product ID*	Biological process
				Low death rate (total n=5)	Fast death rate (n=4)		
7	hypothetical protein	0.000558	accessory gene	5	1	WP_001317782; toxin Ldr, type I toxin-antitoxin system protein	GO:0012501 - programmed cell death
9	*zraS*	0.000414	missense variant	1	4	EG10008; ZraS sensory histidine kinase; cellular response to lead/zink	GO:0071294 - cellular response to zinc ion GO:0071284 - cellular response to lead ion
13	hypothetical protein	0.000363	accessory gene	4	0	EDV82299.1	
17	*ltaE*	0.000331	missense variant	1	4	G6455; LtaE [component of low-specificity L-threonine aldolase]	GO:0006567 - threonine catabolic process GO:0006520 - cellular amino acid metabolic process GO:0006545 - glycine biosynthetic process

Table 14. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype D (continued).

Rank in top 100 RF hits	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of SNP by survival class		Gene/product ID*	Biological process
				Low death rate (total n=5)	Fast death rate (n=4)		
24	Prevent host death protein2C Phd antitoxin	0.000312	accessory gene	4	0	Phd-Doc toxin-antitoxin systems	GO:0012501 - programmed cell death
27	*frmB*	0.00031	missense variant	1	4	G6208; S-formylglutathione hydrolase	GO:0046292 - formaldehyde metabolic process
38	*abgB*	0.000272	missense variant	1	4	G6669; p-aminobenzoyl-glutamate hydrolase subunit B	GO:0046657 - folic acid catabolic process
41	Transposase and inactivated derivatives	0.000265	accessory gene	4	0	WP_001561476.1	
46	ECUMN_0533	0.000243	missense variant	1	4	putative adhesin/invasin-like protein	
47	FIG00640785: hypothetical protein	0.00024	accessory gene	5	1	EDU34564	
49	hypothetical protein	0.000239	accessory gene	3	0	KDA73278	
51	hypothetical protein	0.000232	accessory gene	3	0	EEC25939	
57	*phoA*	0.000222	missense variant	0	3	EG10727; component of alkaline phosphatase	GO:0055114 - oxidation-reduction process

Table 14. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype D (continued).

Rank in top 100 RF hits	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of SNP by survival class		Gene/product ID*	Biological process
				Low death rate (total n=5)	Fast death rate (n=4)		
61	ECUMN_2565	0.000218	missense variant	2	4	putative large extracellular alpha-helical protein	
68	Pertactin precursor	0.000204	accessory gene	5	1	WP_000444402; Autotransporter protein. Virulence factor	
72	hypothetical protein Z4912	0.000185	accessory gene	5	1	KDX51457	
82	*mnmC*	0.000156	missense variant	0	3	G7199; fused 5-methylaminomethyl-2-thiouridine-forming methyltransferase and FAD-dependent demodification enzyme	GO:0002097 - tRNA wobble base modification
90	Phage capsid and scaffold	0.000146	accessory gene	2	4	WP_000123273	
93	*ydfR*	0.00012	missense variant	0	2	G6828; Qin prophage; predicted protein	

Table 14. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype D (continued).

Rank in top 100 RF hits	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of SNP by survival class		Gene/product ID*	Biological process
				Low death rate (total n=5)	Fast death rate (n=4)		
97	*cspB*	0.000084	missense variant	0	2	EG12203; Qin prophage; cold shock protein; predicted DNA-binding transcriptional regulator	GO:0009409 - response to cold GO:0006355 - regulation of transcription, DNA-templated GO:0006950 - response to stress

*ID of the homolog with highest identity is showed for accessory hypothetical proteins

In phylotype B1B, RF identified ten important missense variants in core genes, identified six accessory genes and 84 intergenic and synonymous variants (Table 15). Core genome missense variants occurred in genes that could be involved in stress tolerance (*cmr*, *cadC*) or nutrient utilization (*gmhB, lacY, ydhU, entF*). Among accessory genes, I identified homolog of *Salmonella* Ldr persistence-inducing toxin/antitoxin system in low death rate strains (De la Cruz et al. 2013). Also, phage antitermination protein Q (homolog of lambdoid prophage DLP12 antitermination protein) was present in all high death rate strains. This evidence suggests that lower survival can be explained by transition of phages to lytic phase of cycle.

Table 15. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype B.

Rank in top 100	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of variant by survival class		Gene/product ID*	Biological process
				Low death rate (n=4)	Fast death rate (n=5)		
11	*cadC*	0.00095	missense variant	0	5	EG10133; CadC DNA-binding transcriptional activator	GO:0016563 (obsolete) - obsolete transcription activator activity
14	*yobB*	0.00092	missense variant	0	5	G7015; conserved protein	GO:0016810 - hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds GO:0006807 - nitrogen compound metabolic process
17	*nemA*	0.00092	missense variant	4	0	G6890; N-ethylmaleimide reductase, FMN-linked	GO:0006805 - xenobiotic metabolic process GO:0008748 - N-ethylmaleimide reductase activity
19	*lacY*	0.00091	missense variant	4	0	EG10526; lactose / melibiose: H+ symporter LacY	GO:0008643 - carbohydrate transport GO:0005351 - sugar:proton symporter activity

Table 15. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype B (continued).

Rank in top 100	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of variant by survival class		Gene/product ID*	Biological process
				Low death rate (n=4)	Fast death rate (n=5)		
34	*cmr*	0.00088	missense variant	4	0	G6440; multidrug efflux transporter MdfA (multifunctional)	GO:0015307 - drug:proton antiporter activity GO:0030641 - regulation of cellular pH GO:0015386 - potassium:proton antiporter activity GO:0015385 - sodium:proton antiporter activity
39	*ydhU*	0.00087	missense variant	0	5	G6898; predicted cytochrome	GO:0022904 - respiratory electron transport chain GO:0009055 - electron carrier activity
40	corresponds to STY4175 from Accession AL513382: Salmonella typhi CT18	0.00086	accessory gene	4	0	WP_000079941; homolog of Ldr-like toxin gene from type I toxin/antitoxin (TA) system	GO:0012501 - programmed cell death cell

Table 15. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype B (continued).

Rank in top 100	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of variant by survival class		Gene/product ID*	Biological process
				Low death rate (n=4)	Fast death rate (n=5)		
45	*gmhB*	0.00084	missense variant	4	0	EG11736; D,D-heptose 1,7-bisphosphate phosphatase	GO:0009244 - lipopolysaccharide core region biosynthetic process
48	Phage antitermination protein Q	0.00083	accessory gene	0	5	WP_001204791	
59	*yliI*	0.00082	missense variant	4	0	G6437; aldose sugar dehydrogenase	GO:0016901 - oxidoreductase activity, acting on the CH-OH group of donors, quinone or similar compound as acceptor
65	*ydhU*	0.00081	missense variant	0	5	G6898; predicted cytochrome	GO:0022904 - respiratory electron transport chain GO:0009055 - electron carrier activity
80	FIG00641526: hypothetical protein	0.00076	accessory gene	0	5	ESC99965	

Table 15. List of missense SNPs and accessory genes in top 100 most important genome variants in phylotype B (continued).

Rank in top 100	Gene name/RAST annotations	Mean Decrease Accuracy	Type of variant	Presence of variant by survival class		Gene/product ID*	Biological process
				Low death rate (n=4)	Fast death rate (n=5)		
97	ATP:Cob(I)alamin adenosyltransferase	0.00019	accessory gene	0	4	WP_033561004 (EC 2.5.1.17)	GO:0006779 - porphyrin-containing compound biosynthetic process
98	*entF*	0.00017	missense variant	4	1	EG10264; holo [EntF peptidyl-carrier protein] apo-serine activating enzyme	GO:0009239 - enterobactin biosynthetic process
99	FIG00641027: hypothetical protein	0.00017	accessory gene	4	1	WP_001297382	
100	COG2932: Predicted transcriptional regulator	0.00015	accessory gene	0	4	WP_000259987	

* ID of the homolog with highest identity is showed for accessory hypothetical proteins.

Random forest analysis suggest that a small number of genomic variants may be responsible for enhanced survival in the Hudson silt-loam soil. Obtained results show that both accessory genome content and core genome variation confer to adaptive phenotypes of *E. coli*. While a small number of accessory genes were associated with enhanced survival, some genes are known to promote or degrade persistence in adverse environments, suggesting that a small number of gene acquisition events could be contributing strongly to the phenotype.

According to our results, absence of putative persistence degrading factor (phage antitermination protein Q) may strongly contribute to persistence phenotype. Interestingly, in both phylotypes accessory antitoxin genes from toxin-antitoxin systems were identified as important in low death rate phenotype. This suggests that functionally similar accessory gene variation can explain adaptation across different phylotypes.

Genomic variants in phylotype D showed weaker association with phenotype than variants in phylotype B. Mean decrease of accuracy values in random forest showed generally stronger support for important variants in phylotype B than in phylotype D (Tables 3, 4). In phylotype D strains accessory genes variants were more associated with low death rate in Hudson silt-loam soil. Meanwhile, presence of alternative variants in core genome had tendency to be associated with high death rate. Variation in phylotype B did not show any similar tendencies.

Small number of genetic variants required to generate adaptive phenotype is in agreement with the idea that new ecotypes and, thus, species, can originate rapidly (Koeppel et al. 2013; Mallet 2008). Presence of accessory gene variants such as phage genes and genes from toxin-antitoxin systems can support extreme scenarios of survival/adaptation provided by single gene followed by dissemination of beneficial gene across other ecotypes *via* horizontal gene transfer. Differences in sets of identified adaptive genetic variant across phylotypes B and D suggests that adaptation can be achieved convergently as it was observed in experimental evolution of *E. coli* (Tenaillon et al. 2012). In this study, *E. coli* was experimentally evolved to adapt for elevated temperature - 42.2ºC. Very few genetic variants were shared between evolved populations but most of mutations associated with adaptation to elevated temperature occurred in genes that encode parts of the RNA-polymerase complex or in the *rho* factor gene (factor of termination of transcription). In other words, there were two convergent paths to achieve adaptation.

Examination of patterns of genetic variation in *E. coli* associated with extrahost survival led to the idea that polymorphisms in genes involved in metabolism of substances and/or stress response can provide specific traits that incidentally enhance survival in extrahost habitat (van Elsas et al. 2011a). For example, *E. coli* O157:H7 C279 strain persisted better than C278 at 30ºC in loam soil with autoclaved manure slurry (Topp et al. 2003). Topp et al (2003) hypothesized that strain C279 is possibly better in utilizing specific nutrients, which results in better persistence. Further confirmation for the hypothesis that effectiveness of nutrient metabolism is important was described by Franz et al (2011) on different isolates, though. Long-surviving isolates of *E. coli* O157 were capable of oxidizing organic acids (propionic acid, a-ketobutyric acid, and a-hydroxybutyric acid) much faster than the short-surviving strains (Franz et al. 2011). I have identified by random forest analysis important for survival in Hudson silt loam missense variants in genes *ltaE*, *frmB* and *abgB* for phylotype D (Table 14) and yobB, *gmhB*, *lacY*, and *ydhU* for phylotype B (Table 15), which are directly involved in catabolic or anabolic transformation of metabolites or their transport.

Importance of variation has been repeatedly shown for *rpoS* gene, the general stress response sigma factor, for extrahost persistence in *E. coli* (Rozen and Belkin 2001). Poor persistors among *E. coli* O157:H7 in soil (<160 days) tended to carry SNPs, insertion, and deletions in *rpoS* gene compared to long term survivors (>200 days) (van Hoek et al. 2013). On the other hand, comparative genomic hybridization study showed that populations of *E. coli* can't tolerate a lot of variation in conservative *rpoS* gene because this gene is necessary for stress response (Ihssen et al. 2007). Lack of *rpoS* is rare in isolates of *E. coli* from the environment (Snyder et al. 2012). Moreover, recently it was shown that most of the variation in *rpoS* gene arises during laboratory cultivation, suggesting that this gene is conserved in natural populations of *E. coli* (Bleibtreu et al.

2014). According to our results, observed variation in *rpoS* gene wasn't important for soil survival. However, I observed synonymous variation in *rpoS* gene in phylotype D. I have identified missense variants in other putative stress and xeno-/antibiotics tolerance genes in phylotype B: *nemA* (xenobiotic inactivation) *cmr* (multifunctional multidrug efflux transporter), *cadC* (transcription activator of cadaverin biosynthesis genes, genes involved in excretion under low pH and sensing of high concentration of lysine); phylotype D: *zraS* (sensory histidine kinase of the two component signal transduction system ZraS/ZraR involved in response to the lead/zinc).

Prophage genes and mobile genetic elements that compose significant part of accessory genome in bacteria are one of the foremost forces of initial/immediate adaptation in bacteria due to high variability (Cerveau et al. 2011; Koskella and Brockhurst 2014; Penades et al. 2015), virulence genes spreading (Penades et al. 2015) and potentially, dissemination of antibiotic resistance genes (Colomer-Lluch et al. 2011; Quiros et al. 2014). I report that mobile elements derivatives and prophage genes were associated with survival phenotype in both phylotypes (Table 14, 15). Presence of phage antitermination protein Q was associated with high death rate strains in phylotype B. Lower survival in this case is possibly explained by persistence-degrading effect due to transition of prophage to lytic phase in response to environmental stress.

I found that presence of genes from types 1 and 2 toxin-antitoxin(TA) systems was associated with survival of *E. coli* in Hudson silt loam. The toxin/antitoxin systems are primarily known as selfish genetic systems which can cause postsegregational killing of daughter cells that have lost genome region or plasmid containing this system (Van Melderen and De Bast 2009). Toxin is a protein that has deleterious effect on cell functioning. Antitoxin that inhibits toxin can be either antisense RNA to toxin mRNA (type I and III TA systems) or protein (type II, IV, V) (Ramage et al. 2009). It is suggested that TA systems are involved in induction of persistence

under starvation/stress conditions, growth control and restriction of genome deterioration (Van Melderen 2010; Van Melderen and De Bast 2009). According to Yamaguchi and Inouye (2011) *E. coli* K-12 strain contains at least 36 TA systems (Yamaguchi and Inouye 2011). In strains from phylotype D I have identified only 9 complete type 1 and 2 TA systems and various copies of only toxin or antitoxin genes from type 2 and 4 TA systems. Random forest analysis identified that presence of Phd antitoxin and Ldr-like toxin proteins is associated with better survival in phylotype D (Table 14). Four out of five strains with low death rate in phylotype D contained Phd antitoxin gene (from Phd-Doc type 2 TA system). Phd-Doc TA system was derived from bacteriophage P1. Primary molecular function of Doc-toxin is binding with 30S ribosomal subunit, which leads to inhibition of translation and, thus, reduction of cell growth (Yamaguchi et al. 2011). Seven strains (including reference UMN026 and one of strains with high death rate) contained Ldr-like toxin gene which, thus, is also associated with increased survival in Hudson silt loam. Interestingly, another copy Ldr-like toxin gene was presented in all studied strains of phylotype D. In phylotype B I observed 11 complete type 1, 2 and 4 TA systems and copies of only toxin or antitoxin genes from type 2 TA system. One copy of homolog of Ldr-like toxin genes (Table 15) (type I TA system) among studied strains from phylotype B was presented only in strains with low death rate and in reference strain IAI1.

Variation in content of TA systems is involved in persistence induction(Fasani and Savageau 2015), and our study rather suggest that single TA system can explain better survival of studied strains from both phylotypes. TA systems within one organism interact with each other in synergistic and antagonistic ways that leads to variable contribution to growth phenotype. It was suggested that complex interactions between different TA systems under different environmental conditions generate variation in survival and growth phenotypes that in turn can facilitate ecotype

divergence by growth rate and ability to persist under specific conditions(Goeders and Van Melderen 2014).

Conclusions

Our current results together with previous studies show that, possibly, only few genetic variants are required to achieve adaptation. The idea that a small number of genetic variants are required to generate adaptive phenotype is in agree with idea that new ecotypes and, thus, species, can originate in few generations(for example in less than ~110 generation) (Koeppel et al. 2013; Mallet 2008). Accessory gene variants such as phage genes and TA systems may support scenarios of adaptation provided by single gene. Increase of relative abundance of beneficial genetic variant in bacterial population occurs through proliferation of ecotypes with selectively favored genetic variants and *via* horizontal gene transfer (Shapiro et al. 2012). Differences in sets of identified adaptive genetic variant across phylotypes B and D suggests that adaptive phenotype can be achieved convergently. Genomic variants in phylotype D have showed weaker association with phenotype than variants in phylotype B. Mean decrease of accuracy values in random forest have showed generally stronger support for important variants in phylotype B than in phylotype D (Table 14, 15).

REFERENCES

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M et al. 2008. The RAST server: Rapid annotations using subsystems technology. BMC Genomics. 9.

Baty F, Ritz C, Charles S, Brutsche M, Flandrois J-P, Delignette-Muller M-L. 2015. A toolbox for nonlinear regression in R: The package nlstools. Journal of Statistical Software. 66(5):1-21.

Bergholz PW, Noar JD, Buckley DH. 2011. Environmental patterns are imposed on the population structure of *Escherichia coli* after fecal deposition. Applied and Environmental Microbiology. 77(1):211-219.

Berthe T, Ratajczak M, Clermont O, Denamur E, Petit F. 2013. Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. Applied and Environmental Microbiology. 79(15):4684-4693.

Bingle LEH, Bailey CM, Pallen MJ. 2008. Type VI secretion: A beginner's guide. Current Opinion in Microbiology. 11(1):3-8.

Bleibtreu A, Clermont O, Darlu P, Glodt J, Branger C, Picard B, Denamur E. 2014. The *RpoS* gene is predominantly inactivated during laboratory storage and undergoes source-sink evolution in *Escherichia coli* species. Journal of Bacteriology. 196(24):4276-4284.

Blount ZD. 2015. The unexhausted potential of *E. coli*. Elife. 4.

Bobay L-M, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. Proceedings of the National Academy of Sciences of the United States of America. 112(29):8893-8900.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114-2120.

Boon E, Meehan CJ, Whidden C, Wong DHJ, Langille MGI, Beiko RG. 2014. Interactions in the microbiome: Communities of organisms and communities of genes. FEMS Microbiology Reviews. 38(1):90-118.

Boulesteix A-L, Janitza S, Kruppa J, Koenig IR. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery. 2(6):493-507.

Brandl MT. 2006. Fitness of human enteric pathogens on plants and implications for food safety. Annual Review of Phytopathology. 44:367-392.

Breiman L. 2001. Random forests. Machine Learning. 45(1):5-32.

Brennan FP, O'Flaherty V, Kramers G, Grant J, Richards KG. 2010. Long-term persistence and leaching of *Escherichia coli* in temperate maritime soils. Applied and Environmental Microbiology. 76(5):1449-1455.

Cambardella CA, Moorman TB, Novak JM, Parkin TB, Karlen DL, Turco RF, Konopka AE. 1994. Field-scale variability of soil properties in central Iowa soils. Soil Science Society of America Journal. 58(5):1501-1511.

Carlson SM, Cunningham CJ, Westley PAH. 2014. Evolutionary rescue in a changing world. Trends in Ecology & Evolution. 29(9):521-530.

Cerveau N, Leclercq S, Bouchon D, Cordaux R. 2011. Evolutionary dynamics and genomic impact of Prokaryote transposable elements. Evolutionary Biology: Concepts, Biodiversity, Macroevolution and Genome Evolution.291-312.

Chen PE, Shapiro BJ. 2015. The advent of genome-wide association studies for bacteria. Current Opinion in Microbiology. 25:17-24.

Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brussow H. 2004. Phage-host interaction: An ecological perspective. Journal of Bacteriology. 186(12):3677-3686.

Cingolani P, Platts A, Wang LL, Coon M, Tung N, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpEff: SNPs in the genome of *Drosophila melanogaster* strain *w(1118)*; iso-2; iso-3. Fly. 6(2):80-92.

Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Applied and Environmental Microbiology. 66(10):4555-4558.

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(1):58-65.

Cohan FM. 2001. Bacterial species and speciation. Systematic Biology. 50(4):513-524.

Cohan FM. 2005. Periodic selection and ecological diversity in bacteria. Selective Sweep.78-93.

Colomer-Lluch M, Jofre J, Muniesa M. 2011. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. PLOS One. 6(3).

Contreras-Moreira B, Vinuesa P. 2013. Get_homologues, a versatile software package for scalable and robust microbial pangenome analysis. Applied and Environmental Microbiology. 79(24):7696-7701.

Cordero OX, Polz MF. 2014a. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology. 12(4):263-273.

Cordero OX, Polz MF. 2014b. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology. 12(4):263-273.

Crowther TW, Maynard DS, Leff JW, Oldfield EE, McCulley RL, Fierer N, Bradford MA. 2014. Predicting the responsiveness of soil biodiversity to deforestation: A cross-biome study. Global Change Biology. 20(9):2983-2994.

Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. 2013. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clinical Microbiology Reviews. 26(4):822-880.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research. 14(7):1394-1403.

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

De la Cruz MA, Zhao W, Farenc C, Gimenez G, Raoult D, Cambillau C, Gorvel J-P, Meresse S. 2013. A toxin-antitoxin module of *Salmonella* promotes virulence in mice. PLOS Pathogens. 9(12).

de los Angeles Dublan M, Federico Ortiz-Marquez JC, Lett L, Curatti L. 2014. Plant-adapted *Escherichia coli* show increased lettuce colonizing ability, resistance to oxidative stress and chemotactic response. PLOS One. 9(10).

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 43(5):491-+.

Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. Trends in Microbiology. 18(7):315-322.

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.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology. 2(5):414-424.

Dudley EG, Thomson NR, Parkhill J, Morin NP, Nataro JP. 2006. Proteomic and microarray characterization of the *aggr* regulon identifies a pheu pathogenicity island in enteroaggregative *Escherichia coli*. Molecular Microbiology. 61(5):1267-1282.

Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SAFT. 2013. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Briefings in Functional Genomics. 12(4):366-380.

Dutilh BE, Thompson CC, Vicente AC, Marin MA, Lee C, Silva GG, Schmieder R, Andrade BG, Chimetto L, Cuevas D et al. 2014a. Comparative genomics of 274 *Vibrio cholerae* genomes reveals mobile functions structuring three niche dimensions. BMC genomics. 15:654.

Dutilh BE, Thompson CC, Vicente ACP, Marin MA, Lee C, Silva GGZ, Schmieder R, Andrade BGN, Chimetto L, Cuevas D et al. 2014b. Comparative genomics of 274 *Vibrio cholerae* genomes reveals mobile functions structuring three niche dimensions. BMC Genomics. 15.

Engelberg-Kulka H, Hazan R, Amitai S. 2005. Mazef: A chromosomal toxin-antitoxin module that triggers programmed cell death in bacteria. Journal of Cell Science. 118(19):4327-4332.

Fasani RA, Savageau MA. 2015. Unrelated toxin-antitoxin systems cooperate to induce persistence. Journal of the Royal Society Interface. 12(108).

Finkel SE, Kolter R. 2001. DNA as a nutrient novel role for bacterial competence gene homologs. Journal of Bacteriology. 183(21):6288-6293.

Franz E, van Hoek AHAM, Bouw E, Aarts HJM. 2011. Variability of *Escherichia coli* O157 strain survival in manure-amended soil in relation to strain origin, virulence profile, and carbon nutrition profile. Applied and Environmental Microbiology. 77(22):8088-8096.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science. 315(5811):476-480.

Galili T. 2015. Dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 31(22):3718-3720.

Garneau JE, Dupuis M-E, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. 2010. The CRISPR/CAS bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 468(7320):67.

Geeraerd AH, Herremans CH, Van Impe JF. 2000. Structural model requirements to describe microbial inactivation during a mild heat treatment. International Journal of Food Microbiology. 59(3):185-209.

Germain E, Castro-Roa D, Zenkin N, Gerdes K. 2013. Molecular mechanism of bacterial persistence by HipA. Molecular Cell. 52(2):248-254.

Godreuil S, Cohan F, Shah H, Tibayrenc M. 2005. Which species concept for pathogenic bacteria? An E-debate. Infection Genetics and Evolution. 5(4):375-387.

Goeders N, Van Melderen L. 2014. Toxin-antitoxin systems as multilevel interaction systems. Toxins. 6(1):304-324.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Molecular Biology and Evolution. 19(12):2226-2238.

Gonzalez A, Ronce O, Ferriere R, Hochberg ME. 2013. Evolutionary rescue: An emerging focus at the intersection between ecology and evolution. Philosophical Transactions of the Royal Society B-Biological Sciences. 368(1610).

Gordon DM, Bauer S, Johnson JR. 2002. The genetic structure of *Escherichia coli* populations in primary and secondary habitats. Microbiology-SGM. 148:1513-1522.

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

Gupta VK, Chaudhari NM, Iskepalli S, Dutta C. 2015. Divergences in gene repertoire among the reference *Prevotella* genomes derived from distinct body sites of human. BMC Genomics. 16.

Haiko J, Westerlund-Wikstrom B. 2013. The role of the bacterial flagellum in adhesion and virulence. Biology. 2(4):1242-1267.

Haridas S, Breuill C, Bohlmann J, Hsiang T. 2011. A biologist's guide to *de novo* genome assembly using next-generation sequence data: A test with fungal genomes. Journal of Microbiological Methods. 86(3):368-375.

Hazan R, Engelberg-Kulka H. 2004. *Escherichia coli* MazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1. Molecular Genetics and Genomics. 272(2):227-234.

Hendrickson H. 2009. Order and disorder during *Escherichia coli* divergence. PLOS Genetics. 5(1).

Heuer H, Smalla K. 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. FEMS Microbiology Reviews. 36(6):1083-1104.

Ihssen J, Grasselli E, Bassin C, Francois P, Piffaretti J-C, Koester W, Schrenzel J, Egli T. 2007. Comparative genomic hybridization and physiological characterization of environmental isolates indicate that significant (eco-)physiological properties are highly conserved in the species *Escherichia coli*. Microbiology-SGM. 153:2052-2066.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics. 44(2):226-232.

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(7).

Koeppel AF, Wertheim JO, Barone L, Gentile N, Krizanc D, Cohan FM. 2013. Speedy speciation in a bacterial microcosm: New species can arise as frequently as adaptations within a species. ISME Journal. 7(6):1080-1091.

Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B-Biological Sciences. 361(1475):1929-1940.

Konstantinidis KT, Tiedje JM. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: Advancements and challenges ahead. Current Opinion in Microbiology. 10(5):504-509.

Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: A review. Plant Methods. 9.

Koskella B, Brockhurst MA. 2014. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiology Reviews. 38(5):916-931.

Kovalenko KE, Thomaz SM, Warfe DM. 2012. Habitat complexity: Approaches and future directions. Hydrobiologia. 685(1):1-17.

Kung VL, Ozer EA, Hauser AR. 2010. The accessory genome of *Pseudomonas aeruginosa*. Microbiology and Molecular Biology Reviews. 74(4):621-641.

Kuramae EE, Yergeau E, Wong LC, Pijl AS, van Veen JA, Kowalchuk GA. 2012. Soil characteristics more strongly influence soil bacterial communities than land-use type. FEMS Microbiology Ecology. 79(1):12-24.

Kyle JE, Ferris FG. 2013. Geochemistry of virus-prokaryote interactions in freshwater and acid mine drainage environments, Ontario, Canada. Geomicrobiology Journal. 30(9):769-778.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754-1760.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Research. 13(9):2178-2189.

Liu X, Han S, Wang Z, Gelernter J, Yang B-Z. 2013. Variant callers for next-generation sequencing data: A comparison study. PLOS One. 8(9).

Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences of the United States of America. 104(27):11436-11440.

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. 2004. Screening large-scale association study data: Exploiting interactions using random forests. BMC Genetics. 5.

Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research. 21(6):936-939.

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 of the United States of America. 108(17):7200-7205.

Maddamsetti R, Hatcher PJ, Cruveiller S, Medigue C, Barrick JE, Lenski RE. 2015. Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict where synonymous substitutions occur in a long-term experiment. Molecular Biology and Evolution. 32(11):2897-2904.

Mallet J. 2008. Hybridization, ecological races and the nature of species: Empirical evidence for the ease of speciation. Philosophical Transactions of the Royal Society B-Biological Sciences. 363(1506):2971-2986.

Mallon CA, van Elsas JD, Salles JF. 2015. Microbial invasions: The process, patterns, and mechanisms. Trends in Microbiology. 23(11):719-729.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR et al. 2006. Microbial biogeography: Putting microorganisms on the map. Nature Reviews Microbiology. 4(2):102-112.

Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, Rokas A, Yandava CN, Engels R, Zeng E et al. 2008. Dynamics of *Pseudomonas aeruginosa* genome evolution. Proceedings of the National Academy of Sciences of the United States of America. 105(8):3100-3105.

Maurice CF, Bouvier C, de Wit R, Bouvier T. 2013. Linking the lytic and lysogenic bacteriophage cycles to environmental conditions, host physiology and their variability in coastal lagoons. Environmental Microbiology. 15(9):2463-2475.

Mayden RL. 1997. A hierarchy of species concepts: The denouement in the saga of the species problem. Systematics Association Special Volume Series; Species: The units of biodiversity. 54:381-424.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. Genome Research. 20(9):1297-1303.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Current Opinion in Genetics & Development. 15(6):589-594.

Melendrez MC, Becraft ED, Wood JM, Olsen MT, Bryant DA, Heidelberg JF, Rusch DB, Cohan FM, Ward DM. 2016. Recombination does not hinder formation or detection of ecological species of *Synechococcus* inhabiting a hot spring cyanobacterial mat. Frontiers in Microbiology. 6.

Mendes R, Garbeva P, Raaijmakers JM. 2013. The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiology Reviews. 37(5):634-663.

Meric G, Kemsley EK, Falush D, Saggers EJ, Lucchini S. 2013. Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli*. Environmental Microbiology. 15(2):487-501.

Naomi S-I. 2011. On the integrated frameworks of species concepts: Mayden's hierarchy of species concepts and de queiroz's unified concept of species. Journal of Zoological Systematics and Evolutionary Research. 49(3):177-184.

Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P et al. 2013. Patterns and processes of microbial community assembly. Microbiology and Molecular Biology Reviews. 77(3):342-356.

Niehus R, Mitri S, Fletcher AG, Foster KR. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nature Communications. 6.

Ochman H, Wilson AC. 1987. Evolution in bacteria - evidence for a universal substitution rate in cellular genomes. Journal of Molecular Evolution. 26(1-2):74-86.

Ochoa-Hueso R, Rocha I, Stevens CJ, Manrique E, Jose Lucianez M. 2014. Simulated nitrogen deposition affects soil fauna from a semiarid mediterranean ecosystem in central Spain. Biology and Fertility of Soils. 50(1):191-196.

Orsi RH, Stoppe NC, Sato MI, Ottoboni LM. 2007. Identification of *Escherichia coli* from groups A, B1, B2 and D in drinking water in Brazil. Journal of water and health. 5(2):323-327.

Paradis E, Claude J, Strimmer K. 2004. Ape: Analyses of phylogenetics and evolution in R language. Bioinformatics. 20(2):289-290.

Penades JR, Chen J, Quiles-Puchalt N, Carpena N, Novick RP. 2015. Bacteriophage-mediated spread of bacterial virulence genes. Current Opinion in Microbiology. 23:171-178.

Pitout JDD. 2012. Extraintestinal pathogenic *Escherichia coli*: A combination of virulence with antibiotic resistance. Frontiers in Microbiology. 3.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: The causes and consequences of codon bias. Nature Reviews Genetics. 12(1):32-42.

Polz MF, Hunt DE, Preheim SP, Weinreich DM. 2006. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. Philosophical Transactions of the Royal Society B-Biological Sciences. 361(1475):2009-2021.

Puigbo P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. BMC Biology. 12.

Puntoni V. 1952. Sixth international congress of microbiology. Schweizerische Zeitschrift fur Pathologie und Bakteriologie Revue suisse de pathologie et de bacteriologie. 15(4):538-538.

Quiros P, Colomer-Lluch M, Martinez-Castillo A, Miro E, Argente M, Jofre J, Navarro F, Muniesa M. 2014. Antibiotic resistance genes in the bacteriophage DNA fraction of human fecal samples. Antimicrobial Agents and Chemotherapy. 58(1):606-609.

Ramage HR, Connolly LE, Cox JS. 2009. Comprehensive functional analysis of *Mycobacterium tuberculos* is toxin-antitoxin systems: Implications for pathogenesis, stress responses, and evolution. PLOS Genetics. 5(12).

Ratajczak M, Laroche E, Berthe T, Clermont O, Pawlak B, Denamur E, Petit F. 2010. Influence of hydrological conditions on the *Escherichia coli* population structure in the water of a creek on a rural watershed. BMC microbiology. 10:222.

Read TD, Massey RC. 2014. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: A new direction for bacteriology. Genome Medicine. 6.

Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009. Biogeography of the *Sulfolobus islandicus* pan-genome. Proceedings of the National Academy of Sciences of the United States of America. 106(21):8605-8610.

Ricciardi-Rigault M, Bird DF, Prairie YT. 2000. Changes in sediment viral and bacterial abundances with hypolimnetic oxygen depletion in a shallow eutrophic Lac Brome (Quebec, Canada). Canadian Journal of Fisheries and Aquatic Sciences. 57(6):1284-1290.

Rossello-Mora R, Amann R. 2015. Past and future species definitions for Bacteria and Archaea. Systematic and Applied Microbiology. 38(4):209-216.

Rozen Y, Belkin S. 2001. Survival of enteric bacteria in seawater. FEMS Microbiology Reviews. 25(5):513-529.

Russell AB, Peterson SB, Mougous JD. 2014. Type VI secretion system effectors: Poisons with a purpose. Nature Reviews Microbiology. 12(2):137-148.

Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. 2015. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. Genome Research. 25(1):119-128.

Santo Domingo JW, Bambic DG, Edge TA, Wuertz S. 2007. Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. Water Res. 41(16):3539-3552.

Savageau MA. 1983. *Escherichia coli* habitats, cell-types, and molecular mechanisms of gene-control. American Naturalist. 122(6):732-744.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science. 336(6077):48-51.

Shapiro BJ, Polz MF. 2015. Microbial speciation. Cold Spring Harbor perspectives in biology. 7(10):a018143-a018143.

Sharma R, Polkade AV, Shouche YS. 2015. 'Species concept' in microbial taxonomy and systematics. Current Science. 108(10):1804-1814.

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J, Falush D. 2013. Genome-wide association study identifies vitamin B-5

biosynthesis as a host specificity factor in *Campylobacter*. Proceedings of the National Academy of Sciences of the United States of America. 110(29):11923-11927.

Sikorski J. 2015. The prokaryotic biology of soil. Soil Organisms. 87(1):1-28.

Snyder E, Gordon DM, Stoebel DM. 2012. *Escherichia coli* lacking *RpoS* are rare in natural populations of non-pathogens. G3-Genes Genomes Genetics. 2(11):1341-1344.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: Building the web of life. Nature Reviews Genetics. 16(8):472-482.

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kampfer P, Maiden MCJ, Nesme X, Rossello-Mora R, Swings J, Truper HG et al. 2002. Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology. 52:1043-1047.

Staley JT, Konopka A. 1985. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review of Microbiology. 39:321-346.

Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, Rockhold ML, Konopka A. 2013. Quantifying community assembly processes and identifying features that impose them. ISME Journal. 7(11):2069-2079.

Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Koenig IR, Zhang H, Sun YV. 2009. Machine learning in genome-wide association studies. Genetic Epidemiology. 33:S51-S57.

Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The molecular diversity of adaptive convergence. Science. 335(6067):457-461.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. Nature Reviews Microbiology. 8(3):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 of the United States of America. 102(39):13950-13955.

Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: The bacterial pan-genome. Current Opinion in Microbiology. 11(5):472-477.

Texier S, Prigent-Combaret C, Gourdon MH, Poirier MA, Faivre P, Dorioz JM, Poulenard J, Jocteur-Monrozier L, Moenne-Loccoz Y, Trevisan D. 2008. Persistence of culturable *Escherichia coli* fecal contaminants in dairy alpine grassland soils. Journal of Environmental Quality. 37(6):2299-2310.

Topp E, Welsh M, Tien YC, Dang A, Lazarovits G, Conn K, Zhu H. 2003. Strain-dependent variability in growth and survival of *Escherichia coli* in agricultural soil. FEMS Microbiology Ecology. 44(3):303-308.

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(1).

Tymensen LD, Pyrdok F, Coles D, Koning W, McAllister TA, Jokinen CC, Dowd SE, Neumann NF. 2015. Comparative accessory gene fingerprinting of surface water *Escherichia coli* reveals genetically diverse naturalized population. Journal of Applied Microbiology. 119(1):263-277.

van Elsas JD, Semenov AV, Costa R, Trevors JT. 2011a. Survival of *Escherichia coli* in the environment: Fundamental and public health aspects. ISME Journal. 5(2):173-183.

van Elsas JD, Semenov AV, Costa R, Trevors JT. 2011b. Survival of *Escherichia coli* in the environment: Fundamental and public health aspects. The ISME journal. 5(2):173-183.

van Hoek AHAM, Aarts HJM, El B, van Overbeek WM, Franz E. 2013. The role of RpoS in *Escherichia coli* O157 manure-amended soil survival and distribution of allelic variations among bovine, food and clinical isolates. FEMS Microbiology Letters. 338(1):18-23.

Van Melderen L. 2010. Toxin-antitoxin systems: Why so many, what for? Current Opinion in Microbiology. 13(6):781-785.

Van Melderen L, De Bast MS. 2009. Bacterial toxin-antitoxin systems: More than selfish entities? PLOS Genetics. 5(3).

Velappan N, Sblattero D, Chasteen L, Pavlik P, Bradbury ARM. 2007. Plasmid incompatibility: More compatible than previously thought? Protein Engineering Design & Selection. 20(7):309-313.

Vellend M. 2010. Conceptual synthesis in community ecology. Quarterly Review of Biology. 85(2):183-206.

Verhougstraete MP, Martin SL, Kendall AD, Hyndman DW, Rose JB. 2015. Linking fecal bacteria in rivers to landscape, geochemical, and hydrologic factors and sources at the basin scale. Proceedings of the National Academy of Sciences of the United States of America. 112(33):10419-10424.

Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. Current Opinion in Microbiology. 23:148-154.

Vulic M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proceedings of the National Academy of Sciences of the United States of America. 94(18):9763-9767.

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009a. Cryptic lineages of the genus *Escherichia*. Applied and Environmental Microbiology. 75(20):6534-6544.

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009b. Cryptic lineages of the genus *Escherichia*. Applied and Environmental Microbiology. 75(20):6534-6544.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences of the United States of America. 95(12):6578-6583.

Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiology Reviews. 35(5):957-976.

Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW. 2007. Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biology. 8(12).

Winfield MD, Groisman EA. 2003. Role of nonhost environments in the lifestyles of *Salmonella* and *Escherichia coli*. Applied and Environmental Microbiology. 69(7):3687-3694.

Wirth T, Falush D, Lan RT, 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(5):1136-1151.

Yamaguchi Y, Inouye M. 2011. Regulation of growth and death in *Escherichia coli* by toxin-antitoxin systems. Nature Reviews Microbiology. 9(11):779-790.

Yamaguchi Y, Park J-H, Inouye M. 2011. Toxin-antitoxin systems in Bacteria and Archaea. Annual Review Genetics, Vol 45. 45:61-79.

Yarza P, Yilmaz P, Pruesse E, Gloeckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzeby J, Amann R, Rossello-Mora R. 2014. Uniting the classification of cultured and uncultured

bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology. 12(9):635-645.

Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. 2014. Performance comparison of SNP detection tools with Illumina exome sequencing data-an assessment using both family pedigree information and sample-matched SNP array data. Nucleic Acids Research. 42(12).

Zerbino DR, Birney E. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research. 18(5):821-829.

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

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: A fast phage search tool. Nucleic Acids Research. 39:W347-W352.

Ziegler A, Koenig IR, Thompson JR. 2008. Biostatistical aspects of genome-wide association studies. Biometrical Journal. 50(1):8-28.