Molecular and ecological characterisation of *Escherichia coli* from plants

Guillaume Méric

A thesis submitted to the **University of East Anglia** (Norwich, UK)
in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

IN BIOLOGY

September 2011

Institute of Food Research
Norwich Research Park
Norwich, NR4 7UA
United Kingdom

Abstract

Escherichia coli is routinely isolated from vegetables and there is increasing evidence that plants are a secondary reservoir for commensal and pathogenic strains, but the ecological factors involved in the persistence of *E. coli* on plants are not clear. In this thesis, a comparative study was undertaken combining phenotypic and phylogenetic analyses of *E. coli* isolates from salads grown in the UK and the faeces of mammalian hosts. *In vitro* phenotypic profiling revealed significant differences according to the source of isolation: strains from plants were in the majority from phylogroup B1, displayed lower siderophore production, greater motility, higher biofilm production, and better growth on the aromatic compounds and sucrose. However, plant-associated isolates reached lower growth yields on many carbon sources, including several amino acids and common carbohydrates such as glucose and mannitol. The data obtained indicate that in addition to lateral gene transfer, variation (regulation or uptake) in core metabolic functions plays an important role in *E. coli* ecological adaptation. When the discriminating phenotypes were combined to generate a plant association index (PAi) to rank strains according to their potential to persist on plants, a strong association between PAi and phylogeny was found, notably high levels in phylogroup B1 and low levels in phylogroup B2 which could potentially constitute a good predictor for host specialisation and generalisation in *E. coli.* As a more applied and preliminary investigation, the question of how a strain with a medium level of PAi (GMB30) can influence the resident microflora of field- and laboratory-grown spinach was also addressed. Overall, this study shows that despite frequent acquisition and loss of traits associated with nonhost environments, the *E. coli* phylogroups differ substantially in their transmission ecology, and in the adaptation levels to their host.

Author’s declaration

I hereby certify that the work contained in this thesis is entirely the result of my own work, except where reference is made to other authors. It has not been submitted in any other form to the University of East Anglia or any other University.

Dr E. Katherine Kemsley (IFR, Norwich, UK) is credited for producing statistical analyses requiring the MatLab software and presented in this thesis, where mentioned.

Dr Stephanie Schüller (IFR / UEA School of Medicine, Norwich, UK) has handled and performed a metabolic profiling experiment of the O104:H4 strain of *E. coli* described in this thesis.

Materials from Chapter 2 to 4 are planned to be included in one or two scientific publications.

Guillaume Méric

Acknowledgements

My most sincere acknowledgements go to the people cited below, for their help and support, and without whom this part of my life would simply not have been spent in this most exciting and constructing way:

Sacha Lucchini, for accepting to get on board of this story as it was starting. His scientific input and feedback on the work presented here were decisive, and I am very grateful for his patience, his precious time and the numerous stimulating conversations we exchanged, I hope this will continue!;

Tim Brocklehurst and Elizabeth Saggers for their expertise, support and help, as well as for their trust; Sue for her advice and support;

Kate Kemsley from the Bioinformatics & Statistics department at IFR; her competent work and invaluable input were crucial in the science we did together and the work presented here; may our (hopefully!) joint publication be a nice reminder of this enjoyable collaboration;

Daniel Falush, Xavier Didelot and Samuel Sheppard who gave very valuable intellectual input and insights on the work in the last few months;

Pani Tourlomousis and Richard Bailey for their help, friendly advice and guidance of on the DGGE work;

Caroline Weight, Ida Porcelli, Isabelle Hautefort, Regis Stentz, Mark Reuter *et al.* for many others things;

Additionally, I thank my friends and family in France, here at IFR or across the road at the Sainsbury Laboratory, who made these occasionally stressful times worth enjoying, especially to Kee Hoon, Cécile, Milena, Christiaan, Chris, Simon “Mitch”, Joo Yong the Dragon, Cam, Popo, Pierre and others. A mes parents, ma sœur Mathilde et le reste de ma famille, merci mille fois de m’avoir soutenu et aidé pendant toutes ces années d’études !

Finally and most importantly, thank you Sophie, for being here. Your loving support makes everything in my life enjoyable and worthwhile.

List of contents

1. Introduction

1.1. Evolutionary biology and population genetics of *E. coli*

1.1.1. Generalities on *E. coli*

1.1.1.1. Brief historical background

At the times of Robert Koch (1843-1910), first theoretician of the link between diseases and pathogenic bacteria, first observer (or rediscoverer) of *Bacillus anthracis* (1877), *Mycobacterium tuberculosis* (1882), *Vibrio cholerae* (1883) and Nobel Prize laureate in 1905, bacteriology was a controversial field of science. The popular "miasmatic theory" suggested that diseases such as cholera were caused by abiotic "poisonous vapour" in the air (or "miasma") and its subsequent inhalation by humans. Belief in the miasmatic hypothesis, despite having been disproved by John Snow during the 1854 cholera epidemic of London, was widespread among physicians even at the end of the 19th century. Koch's work was routinely criticised by then-influential physicians. The debate was so virulent at the time that Max von Pettenkofer, a renowned professor in Munich, personally drank a suspension of *V. cholerae* which he had received from Koch (who had diluted it), thus contracting a weak dysentery (Friedmann, 2006). More and more accumulation of evidence gradually came to validate the "germ theory of disease" at the beginning of the 20th century.

A brilliant yet nowadays rather publicly unknown Bavarian physician called Theodor Escherich (1957-1911) (**Figure 1.1**) played a considerable role in the acceptance of the germ theory. As a peadiatrician, Escherich's main focus was to understand the

cause of then-devastating neonatal infections. Indeed in the early 1900s, a mortality rate of 80% at birth was not uncommon in Europe and child mortality remained around 20% until one year of age. As a consequence, newborn and infants were rarely admitted to hospitals before one year old. Physicians were usually inefficient and powerless against bacterial infections, and hygienic practices were not widespread in hospitals.

Figure 1.1. Theodor Escherich (1857-1911) pictured around 1900. (Shulman et al., 2007) This picture is in the public domain.

In 1885, after having learnt the basics of pure culturing and sterile manipulation from one of Koch's students, Escherich compared the microflora of both meconium and feaces of newborns at different stages after birth. He observed and isolated a variety of organisms from healthy and sick patients, including a rod-shaped bacillus he named *Bacterium coli commune* (*coli*, from Latin "of the colon") which was present in great abundance in all faecal samples he tested. He predicted that *B. coli commune* was an inhabitant of the lower parts of the gastrointestinal tract and used the staining

techniques recently developed by Christian Gram. Escherich was also among the first to develop anaerobic culture methods. He demonstrated that under anaerobic conditions, the growth of some bacteria was solely dependent on carbohydrate fermentation, of which he identified the produced gas (Shulman et al., 2007). Investigating the pathogenic properties of the newly found bacteria, Escherich injected various suspensions to animals, successfully causing disease in guinea pigs and cats. In 1895, he also postulated that *B. coli commune* was responsible for bladder infections by isolating unusually large quantities of it from the urine of symptomatic young girls.

There is a strong belief that *E. coli* became an interest of research only with the development of molecular biology in the 1940s, yet *B. coli commune* was already an intensive topic of study in bacteriology laboratories across the world right after its discovery, because of its ease of isolation and cultivation and its very short generation time, rather than of general interest for faecal microflora. A good summary of pre-molecular biology era studies on *E. coli* can be found elsewhere (Friedmann, 2006). It is worth noting that Escherich is also probably the discoverer of another major bacterial pathogen, *Campylobacter jejuni*, after the 1884 Naples epidemic (Kist, 1986). Theodor Escherich died of a heart stroke at 53 years-old in 1911, weakened by the death of his youngest son from appendicitis. He had led a brilliant career as a socio-paediatrician, teaching professor and children's hospital director. Escherich is considered as one of the very first paediatric physician with an interest in infectious diseases (Shulman et al., 2007).

The next year in 1912, two physicians specialised in tropical medicine, Aldo Castellani and Albert J. Chalmers suggested *Bacterium coli commune* to be renamed to *Escherichia coli* in honour of Theodor Escherich. However, the appellation *Bacterium coli* persisted in the literature for a few decades (Jacques Monod himself did not refer even a single time to the genus *Escherichia* in his 1942 doctoral thesis, preferring the genus *Bacterium*). In the 1930s, a clear distinction was commonly accepted between bacteria and eukaryotes in the sense that bacteria were evolving and mutating faster, which often lead to the use of the *Escherichia coli mutabile* nomenclature. The apparition of authoritative studies in the 1940s on the genetics of laboratory strains like *E. coli* K-12 and B probably contributed to the generalisation of the use of the *Escherichia* genus over *Bacterium*.

The genus name *Escherichia* was officially introduced by the Judicial Commission of the International Committee on Bacteriological Nomenclature in 1958. As of 2011, the taxonomy based on molecular methods places the genus *Escherichia* and the species name *coli* in the Proteobacteria phylum (γ-Proteobacteria class), in the *Enterobacteriaceae* family within the *Enterobacteriales* order. Its full Linnaean name is "*Escherichia coli* (Migula 1895) Castellani and Chalmers 1919". Despite this clear and accepted taxonomy, it is not trivial to reconstruct accurate detailed phylogenies and infer correct ancestries for *E. coli*, *Enterobacteriaceae*, and more generally among bacteria, as illustrated in section 1.1.2. In the next section, a brief introduction to other *Enterobacteriaceae* and *Escherichia sp.* is provided.

1.1.1.2. E. coli and other *Enterobacteriaceae*

Members of the *Enterobacteriaceae* family show common morphological features, as they all are typically 1 to 5µm-long Gram-negative bacilli, facultative anaerobes that do not produce spores. There is however a high ecological versatility among them as many live in the gastrointestinal tract (GIT) of animals, but also in soil, water or sediment. *Enterobacteriaceae* are generally associated with the animal gastrointestinal tract, with some strains in genera *Escherichia*, *Salmonella*, *Klebsiella*, *Serratia*, *Citrobacter* or *Yersinia* being important animal pathogens. While the name itself can be misleading, *Enterobacteriaceae* also comprise plant-associated bacteria and phytopathogens, such as members of the genera *Erwinia*, *Pectobacterium*, *Dickeya*, *Enterobacter*, *Brenneria* or *Pantoea*. *Enterobacteriaceae* are within the γ-Proteobacteria class, itself close to the β-Proteobacteria class (Wu et al., 2009) which is notably composed of the pathogenic families *Neisseriales* and *Burkholderiales*.

Over the last century, many species have been described and added in the *Escherichia* genus. However, after reclassification of some of them based on their genetic dissimilarity (notably *E. blattae*, *E. hermanii* and *E. vulneris*), it is now considered that there are only 3 distinct species in the *Escherichia* genus: *E. coli*, *E. albertii* and *E. fergusonii*. More recently, it has been observed that some strains initially identified as *E. coli* were phylogenetically distant from the majority of *E. coli* strains although phenotypically indistinguishable from them (Walk et al., 2009). These strains were initially thought to be hybrids (Walk et al., 2007) or ancestral variants of *E. coli* that survived a possible selective sweep (Wirth et al., 2006) but more recent studies suggest that they are not strictly the *coli* species but members of five distinct

Escherichia sp. cryptic lineages which likely represent nascent evolutionary lineages (Walk et al., 2009). By examining current strain collections based mostly on faecal samples, these strains appear in clear minority, indicating that their ecologies or their environmental abundance is probably different from *E. coli* (Clermont et al., 2011; Luo et al., 2011).

Additionally, it is interesting to note that *Shigella* strains, initially based on their ability to cause disease, have been wrongly classified as a separate genus from *Escherichia* (Lan and Reeves, 2002). Since the early observation by Salvador Luria in 1957 that the conjugation frequencies between *Shigella flexneri* and *E. coli* strains were similar to the frequencies within *E. coli* strains (Luria and Burrous, 1957), it has been observed that *Shigella* strains were phylogenetically part of *E. coli*, either using multilocus enzyme electrophoresis (MLEE) (Ochman et al., 1983) or more recent methods (Touchon et al., 2009; Sims and Kim, 2011). *Shigella* is a good example of the occasional inertia in medical terminology, as this appellation is still confusingly widely used in clinical setting, and thus medical research. The correct classification is to consider *Shigella* as nothing more than a pathotype of *E. coli*, with strong similarities to enteroinvasive *E. coli* (EIEC) and a specific evolutionary history. More details are provided in section 1.2.1 of this thesis.

Comparative genomics have shown that *E. coli* (and enterobacteria) has an "open pan-genome", meaning that the analyses of new genome sequences of *E. coli* increases the number of genes associated with the species. In other words, the diversity of *E. coli* genes is increasing. It is generally observed that an open pan-genome is associated with species that can colonise multiple environments, and thus

have an increased likelihood of exchanging genes, whereas the contrary is observed for bacterial species living in more specific, isolated niches such as *Mycobacterium tuberculosis* or *Chlamydia trachomatis* (Medini et al., 2005). The consequence of an open pan-genome for *E. coli* is a tremendous diversity at the genotypic and phenotypic levels. In the next sections, we will briefly address these.

1.1.2. Genome dynamics in *E. coli* and other bacteria

1.1.2.1. Core and flexible genome

A logical start in understanding the genomic plasticity in *E. coli* is to know how many genes it has. The chromosome length of *E. coli* has been observed to be 4.5 to 5.5 Mb on average (Bergthorsson and Ochman, 1998). This difference of up to 1 Mb (18% to 23% of the total genome size) indicates that genomic variability is surprisingly high in *E. coli*. Some decades ago, one could have imagined this variability to be caused by the addition of up to 1 Mb of variable genetic content, while most of the genome remained conserved. As a matter of fact, the estimated number of "core" genes, common to all *E. coli* is decreasing as more full genomes are available and examined. An early study comparing 15 genomes of mostly pathogenic *E. coli* found 2,200 genes common to all among a total of 13,000 genes in the pan-genome (Rasko et al., 2008). Touchon and colleagues (2009) revised this estimation by identifying 2,000 genes common to 21 genomes. The latest examination of 61 *E. coli* genomes estimated the core genome to be composed of only 993 genes, within a pan-genome of more than 15,000 genes (Lukjancenko et al., 2010). This estimation suggests that more than 90% of the pan-genome is composed of accessory genes (or the so-called

"flexible" genome), reflecting the high level of strain-specific genetic information in *E. coli*.

Interestingly, the *E. coli* pan-genome has been described as "open", meaning that this species is still evolving mainly by genetic acquisition and diversification (this point is discussed further in the next section). In the light of these observations, it is of prime interest to understand what the genome dynamics in *E. coli* are. In other words, what are the evolutionary forces shaping the acquisition and loss of genetic content, and thus the *E. coli* genomes (**Figure 1.2**).

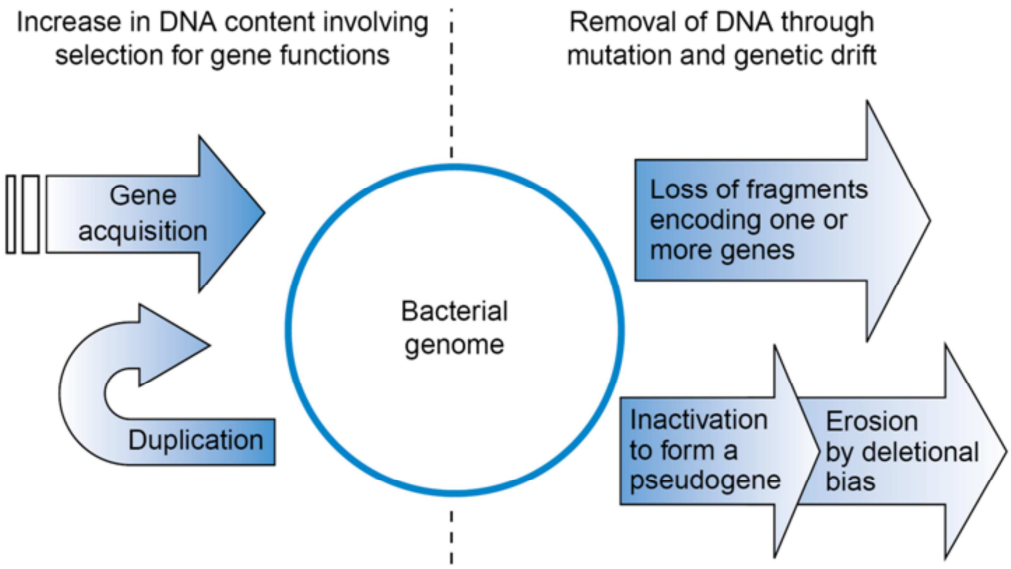


Figure 1.2. Processes involved in genome dynamics (acquisition or loss of genetic content) in bacteria (Mira et al., 2001). This figure is copyrighted by Elsevier Science Ltd.

1.1.2.2. Mechanisms contributing to gene acquisition

As shown in **Figure 1.2**, there are two major ways in which a bacterium can gain additional genetic information. Gene duplication, i.e. the creation of multiple

paralogous copies from an existing gene, has been observed in bacterial and archaeal genomes (Yanai et al., 2000; Jordan et al., 2001; Gevers et al., 2004). Ribosomal genes are a good example of genes that were duplicated before copies diverged (Gevers et al., 2004) and evolve concertedly (Liao, 2000). Concerted evolution is not always observed, and there are many other examples of gene duplication in *E. coli* and other bacteria (Gevers et al., 2004; Serres et al., 2009). However, it seems that the major route of gene acquisition in bacteria is via horizontal gene transfer (HGT), which occurs in 3 distinct ways: by conjugation of exogenous plasmid-borne traits, by transformation, or the natural uptake of DNA fragments secreted by other bacteria (Stentz et al., 2009) or released by dead ones, and finally by transduction of lysogenic bacteriophages into bacterial cells.

DNA acquired from HGT (and therefore called “mobile genetic elements”) can be integrated into the chromosome by homologous recombination or on some occasions kept in a circular form in the bacterial cytoplasm. Upon arrival in a bacterial cell, acquired DNA has different plausible fates directly depending on its degree of homology with sequences present in the recipient genome. When the foreign DNA is not homologous enough, its likelihood of successful homologous recombination decreases while the likelihood of degradation of non-methylated fragments by restriction enzymes increases (Skippington and Ragan, 2011). In other words, a certain degree of sequence homology is required for homologous recombination to work on horizontally acquired DNA (Shen and Huang, 1986; Thomas and Nielsen, 2005). This sequence homology can be observed at different levels, from the strain-specific to the species level and even within larger taxonomic families (Beiko et al., 2005; Toth et al., 2006). Furthermore, it has been suggested that physical proximity

rather than phylogenetic relatedness contributes more to the successful integration of horizontally acquired genes (Matte-Tailliez et al., 2002). Conceivably, this argument does not exclude a certain degree of phylogenetic relatedness too: in a non-stressful environment, the physically close neighbours at the microbial scale are often clones from the same events of binary fission (Didelot and Maiden, 2010). From genomic comparison studies it would, however, seem that although possible, gene transfer between taxonomic families is rarer than between or even within species (Skippington and Ragan, 2011).

The reason why bacteria engage in homologous recombination is still debated (Redfield, 2001; Narra and Ochman, 2006; Michod et al., 2008). As mentioned above, one of the major hypotheses is that homologous recombination is involved in DNA repair and that the incorporation of foreign DNA via HGT is just a way for bacteria to procure templates for the reparation or complementation of damaged DNA (Vos and Didelot, 2009). This function may very well be the most important biological role of homologous recombination (Michod et al., 2008; Didelot and Maiden, 2010) and is conflicting with the "physical proximity only" argument presented above (Matte-Tailliez et al., 2002). The alternative (but not exclusive) "food hypothesis" suggests that the presence of recombinant DNA in bacterial genomes could be a by-product of DNA metabolism, as natural competence and the ability to uptake and use the DNA molecule as a source of nutrients is believed to be an important fitness advantage for competing bacteria (Redfield, 1993; Finkel and Kolter, 2001; Palchevskiy and Finkel, 2006).

The direct impact of HGT on the ecology of *E. coli* is evident, as most of the virulence-associated genes, but also resistance to antibiotics and colonisation-associated factors, are located on mobile genetic elements (sometimes referred as the "mobilome"). Prophages, plasmids, transposons or genomic islands can be maintained and/or integrated in the genome via recombination and spread in multiple strains and species, replicatively or not.

1.1.2.3. Loss of gene content

Two main mechanisms are involved in the loss of genetic content as shown in **Figure 1.2**, either "genetic erosion" (i.e., a gene develops into a truncated, inactivated or degraded version, called a "pseudogene"), or the deletion of genes in a single-step recombinational event, both of which are briefly described below.

Bacteria, unlike eukaryotes, show a linear correlation between the number of protein-coding genes and their genome sizes (Mira et al., 2001). For instance, *Carsonella ruddii* has a genome size of about 160 kb and is an obligate intracellular insect symbiont, with very little metabolic versatility and complexity (Nakabachi et al., 2006). On the other hand, *Pseudomonas aeruginosa* has an average genome size of 6.3 Mb and a great capacity to adapt to multiple environments, without the need for much specialisation (Dobrindt and Hacker, 2001). This robust correlation between genome size and protein-coding genes implies that (a) there is a very limited amount of "junk DNA" in bacterial genomes compared to eukaryotes and (b) that there is a constant evolutionary force leading to the erosion of genes, called the "deletional bias", which is counterbalanced by selection on gene function (Mira et al., 2001). In

other words, all genetic information on bacterial genomes is bound to degrade, unless it is selected for (and assumed useful for the bacterium). The mutational bias towards deletions rather than insertions is also supported by the observation that pseudogenes seem to be eliminated from bacterial genomes more rapidly than the accepted neutral model of stochastic loss, suggesting a possible positively selected mechanism to eliminate non-functional genes, although no molecular mechanism has ever been suggested (Lerat and Ochman, 2004; Kuo and Ochman, 2010). Interestingly, as pseudogenes can only be identified by comparative genomics, the more genomes that are available for a given bacterial species, the more pseudogenes are found (Lerat and Ochman, 2004). Their number has been estimated to be between 80 and 100 in the genome of *E. coli* K-12 strain MG1655 (Ochman and Davalos, 2006; Touchon et al., 2009). It was more recently observed that the number of pseudogenes varied from 45 to 95 in 7 genomes of *E. coli* (Touchon et al., 2009).

There are other mechanisms of genomic rearrangements including inversions, deletions or translocations of large genetic regions (Hughes, 2000). The most common effect of genomic rearrangement is to modify gene synteny, or the order of genes on a chromosome. The comparison of 8 *Yersinia sp.* genome sequences identified no less than 79 genomic inversions (Darling et al., 2008), which indicates that this process is likely to be important in shaping the genomic structure of enterobacteria. Usually, rearrangements occur by recombination between two repeated sequences on the chromosome. One explanation for the existence of these repeats is that they offer a selective advantage in enhancing diversity, generating different sequences at various loci (Ussery et al., 2004). Interestingly, repeats have been lost in endosymbionts (Tamas et al., 2002), suggesting a link between genomic

rearrangements and the ecological need for plasticity to adapt to changing environments. When recombination occurs between two repeats, it can lead to different events: if the two repeats are oriented in opposite direction, recombination will produce an inversion; if they are in the same direction, recombination will produce a circularisation of the sequence included between the two repeats, and this DNA will be lost during cell division, leading to a deletion at this locus.

A good experimental method to observe the global effects of genomic rearrangements is REP-PCR and its derivatives (Versalovic et al., 1993), which are PCR-based fingerprinting methods using primers in the repeat regions. Amplicons obtained by REP-PCR correspond to DNA fragments between repeats. After electrophoresis, a fingerprint showing multiple bands is obtained, and can visually illustrate the diversity of genomic rearrangements if multiple strains are used. This method is used on various *E. coli* isolates in Chapter 3 of this thesis.

1.1.2.4. Impact of recombination on diversity and phylogenetics

The loss of genetic information is not necessarily correlated with the loss of gene content. Gene conversion (i.e. the replacement of one allelic variant by another by recombination without addition or loss of genetic content) is both a divergent and convergent evolutionary force. On one hand, it can lead to an increase in the number of allelic variants in the species but it can also act as a homogenising force, replacing allelic variants by others conferring a higher environmental fitness, which has been observed to be rather prevalent in *E. coli* (Touchon et al., 2009). This mechanism of homogenising gene conversion makes a lot of sense in the context of the selfish gene

theory, for which genes are the basic unit of natural selection rather than organisms or groups of organisms (Werren, 2011). Conceivably, "successful" genes (i.e., genes conferring an increased selective advantage and thus being spread among a large number of bacteria) can predominate within *E. coli* populations after events of lateral transfer and gene conversion.

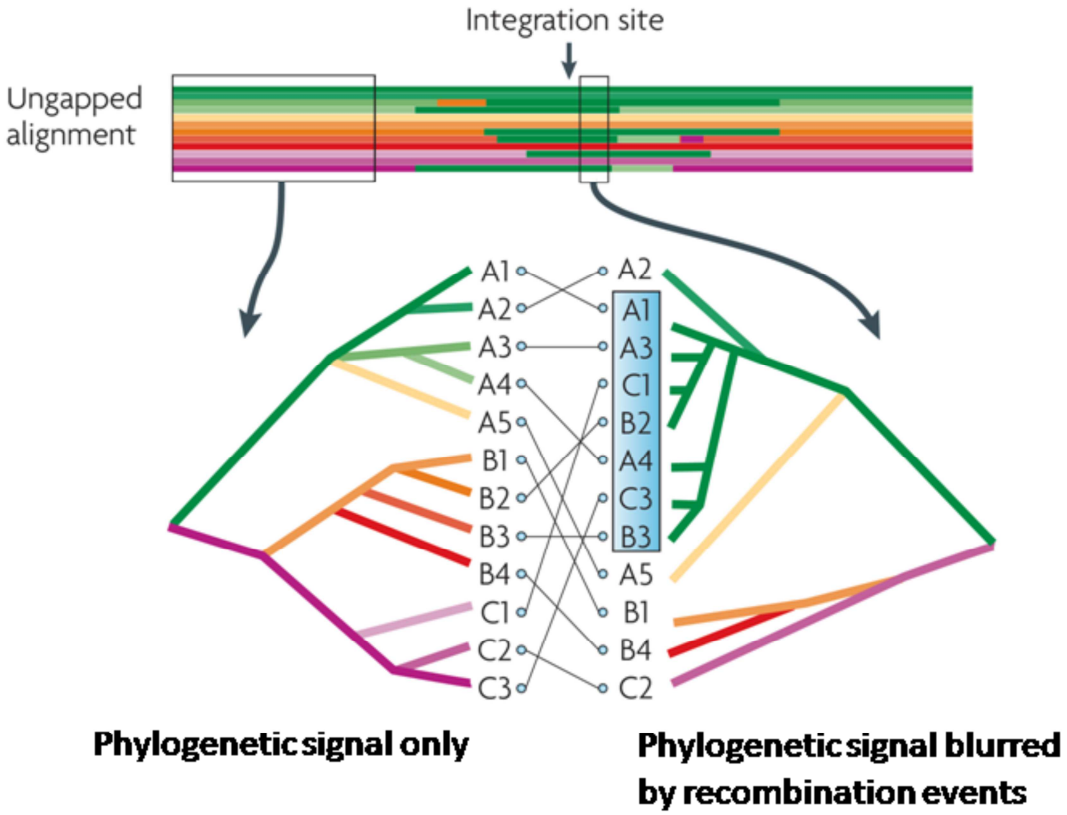


Figure 1.3. Influence of recombination on the reconstruction of phylogenies (Tenaillon et al., 2010). At the top, a sequence alignment between 12 isolates from 3 phylogenetic clades (A, B and C) identifies a recombinational "integration site". Phylogenies based on sequences from this integration site or ungapped alignments without any recombinational effect are incongruent (i.e. inconsistent). Strains from distinct clades appear to be related as an effect of HGT. This figure is copyrighted by MacMillan Publishers Ltd.

This phenomenon can become problematic when reconstructing phylogenies and inferring the ancestry of bacterial strains. Indeed, as illustrated by **Figure 1.3**, if

recombination events are not accounted for, two strains with no shared ancestry would appear to be phylogenetically related on the basis that their difference at the recombined locus is minimal. Conversely, two bacteria that are closely phylogenetically related but diverge at recombined loci can appear to be more distant than they really are (Dykhuizen and Green, 1991). This lack of phylogenetic congruence caused by recombination between different regions of the genome is typical of species with a high level of gene flow such as *E. coli*.

The recombination problem is obviously important when using single genes to reconstruct phylogenies, even if there are some exceptions (Lescat et al., 2009; Sankar et al., 2009). The use of multilocus approaches, such as multilocus sequence typing (MLST) used in this thesis, or even whole genomes (Touchon et al., 2009; Sims and Kim, 2011), can contribute to decrease the phylogenetic signal "blur" caused by recombination. It is indeed unlikely that the same recombination events occurred at multiple loci, providing they are distant enough on the chromosome. In fact, the longer the sequences used for phylogenetic reconstruction are, the more accurate is the produced phylogeny, as the relatively short size of DNA fragments involved in recombination becomes less and less able to interfere with the phylogenetic signal (Tenaillon et al., 2010). For example, if the whole sequence alignment represented in **Figure 1.3** was to be used for phylogenetic reconstruction, as most of it reflects the true phylogenetic history of the strain despite the recombination events, the resulting phylogeny would probably not be far from reality.

This "averaging" strategy consisting of masking the recombination noise by adding more meaningful information works (the minimum corresponding to MLST and the

maximum to a whole-genome approach), but also reflects the difficulty of finding phylogenetic markers that are good representatives of the true phylogenetic history. Even in the high-throughput genomics era, getting large genomic sequences from a large number of strains is still costly for some laboratories, including ours. MLST has long been a method of choice for phylogenetic reconstruction, and a lot of efforts have been made regarding MLST in picking the right combination of genes (called MLST "schemes") that reflect the "purest" phylogenetic signal, if it exists (Spratt, 2004). Providing the right loci are chosen and the right methods are used, the obtained trees can be very similar, or even better to what is obtained using whole genome information (Konstantinidis et al., 2006). Computational methods have also been developed to identify and minimise as much as possible the recombination noise in sequence alignments. Notably, ClonalFrame (Didelot and Falush, 2007) is a method to infer clonal relationships (i.e., genealogy) between bacteria from multilocus data (MLST or whole-genome) by identifying and taking into account the recombination signal (the "clonal frame" refers to the true strain genealogy as if there was no recombination and bacteria could be traced by clonal descent). The major observable difference between a ClonalFrame-based phylogenetic tree and classical ones is that strains with a recombination signal that blurs the phylogenetic signal will be placed equidistant from their common node (i.e., a "multifurcation") whereas classical approaches (such as maximum likelihood or neighbour-joining) will infer a (wrong) relationship regardless of the recombination events (Didelot and Falush, 2007; Didelot and Maiden, 2010). We used ClonalFrame in this thesis to analyse MLST data and described this method more in Chapter 3.

1.1.3. *E. coli* population structure

1.1.3.1. Clonal population structure and phylogenetic groups

The first studies on bacterial population structure were performed about 30 years ago on *E. coli* using multilocus enzyme electrophoresis, or MLEE, which was then mostly used for eukaryotic studies (Selander and Levin, 1980; Selander et al., 1986). This ancestor method of MLST relies on the differential electrophoretic migration of various enzymes (Selander et al., 1986). It has been observed that *E. coli* from various hosts exhibited a relative diversity of their electrophoretic profiles using various enzymes (Selander and Levin, 1980), which were not randomly associated between them, leading to the possibility to define clear groupings of strains (Whittam et al., 1983). In an effort to study further the *E. coli* population structure, a standard reference collection of 72 strains (called "ECOR" for "*E. coli* Reference") was created from a larger set of more than 2600 strains (Ochman and Selander, 1984). The ECOR strains were assembled together on the basis of their diversity of MLEE profiles, possibly encompassing the highest genetic diversity within the *E. coli* species (Ochman and Selander, 1984). Phylogenetic studies based on MLEE profiles first identified 3 major phylogenetic clades of strains (or "phylogroups") called A, B and C (Selander and Levin, 1980). Phylogroup C was later refined into phylogroup D and "ungrouped strains" (Herzer et al., 1990). These ungrouped strains were clustering inconsistently in the early phylogenies and were gathered in the "minor" phylogroup E (Wirth et al., 2006). Their importance was however revised when pathogenic strains of *E. coli* serovar O157:H7 were found to group in clade E, and

subsequent studies have focused on the evolution of EHEC within this group (Wick et al., 2005; Sims and Kim, 2011).

The order in which phylogroups diverged from a common ancestor to all *E. coli* strains has also been debated as different phylogenies indicated different results. It was first thought that phylogroup A was diverging from the others (Herzer et al., 1990) but soon it emerged that either phylogroups B2 and D were basal (Lecointre et al., 1998; Wirth et al., 2006) while phylogroups A and B1 diverged later and are more closely related to each other. Other studies seemed to suggest that B2, rather than D was the first group to diverge (Tenaillon et al., 2010; Sims and Kim, 2011). A whole-genome phylogeny study places D as the most ancient phylogroup, followed by B2 (Touchon et al., 2009).

As mentioned above, early MLEE-based studies observed a high level of linkage disequilibrium in many tested alleles (Selander and Levin, 1980; Whittam et al., 1983; Whittam et al., 1983; Herzer et al., 1990). It was then assumed that recombination in *E. coli* was low, explaining the stability of the groupings and the high linkage disequilibrium observed (i.e., the non-random association of enzymatic profiles, and thus allelic variants) (Selander and Levin, 1980; Hartl and Dykhuizen, 1984). This clonal view of bacterial population structure has been put into question (Maynard Smith et al., 1993) as it has since been shown that *E. coli* strains were actually composed of a very diverse mosaic of mixed ancestry due to recombination (Wirth et al., 2006). In fact, a significant amount of gene flow between the different phylogroups has been detected (Leopold et al., 2011). This phenomenon seems to be a common feature across enteric bacteria, in which phylogenetic incongruence is

observed between phylogenies of orthologous genes and the clonal genealogy of species (Retchless and Lawrence, 2010). However, it would seem that this discordance between enteric bacteria is not linked to ongoing recombination events. Alternatively, authors suggest a fragmented speciation model, defined as the "stepwise acquisition of genetic isolation" upon sudden adaptive changes (Retchless and Lawrence, 2010). Recently, this model has been questioned using strains within the *Escherichia* genus (Luo et al., 2011).

1.1.3.2. Ecological considerations

Whether enteric bacteria evolve by recombination or "fragmented speciation" events, the genetic flow between *E. coli* phylogroups has interestingly been observed to be unbalanced (Leopold et al., 2011). In other words, some phylogroups exchange more genes between them than with others. This could be a reflection of different degrees of phylogenetic relatedness of the different phylogroups but also of different ecological strategies by strains of the different phylogroups. It is possible that the absence of observed gene flow between two groups of strains is simply because the ecology of these two groups leads to different proportions of these strains in various environments, and therefore fewer opportunities for the physical contact required for HGT (Leopold et al., 2011). Indeed, a great variation in the proportions of different phylogroups has been observed in various environmental samples, suggesting that complex ecological factors shape the distribution of phylogenetic groups in different environments. The observation of various phylogroup distributions in various environments could be the consequence of different levels of adaptation of each phylogroup to these various environments.

The link between the phylogeny and different ecologies has been made in humans, in which it has been observed that faecal and urine isolates were preferentially from phylogroup A and B1 (Duriez et al., 2001). However, these figures have been criticised for their sampling bias (Zhang et al., 2002) and it seems that urinary and rectal *E. coli* in humans are in fact more dominated by B2 isolates rather than A and B1 (**Table 1.1**). This trend seems to be confirmed by the fact that B2 are overwhelmingly better intestinal persisters than any other group (Nowrouzian et al., 2005). In this last study, authors observed that 60% of "resident" strains (defined as persisting for more than 3 weeks in the gut, as opposed to "transient" strains) were from phylogroup B2 in Swedish infants (Nowrouzian et al., 2005). Comparatively, phylogroup B1 was the less represented among the resident strains, with only about 5% of strains being resident (**Table 1.1**) (Nowrouzian et al., 2005).

Table 1.1. (next page) Comparison of results from various studies examining the population structure of *E. coli* populations in different environments. The heatmap reflects the level of prevalence in different environments for each phylogroup, with red for high, white for medium and blue for low. The number of strains tested is indicated to contextualise the possible significance of observed percentages.

Source	Type of environment	n (total)	A	B1	B2	D	References
Swedish infants ("resident")	Human host	58	18%	5%	60%	18%	(Nowrouzian et al., 2005)
Swedish infants ("transient")	Human host	19	23%	23%	21%	34%	
Humans from Mali	Human host	55	24%	58%	16%	2%	(Duriez et al., 2001)
Humans from Croatia	Human host	57	35%	32%	14%	19%	
Humans from France	Human host	56	40%	34%	15%	11%	
Humans from Brazil	Human host	94	40%	9%	13%	38%	(Carlos et al., 2010)
Women from Michigan, USA	Human host	181	14%	8%	59%	20%	(Zhang et al., 2002)
Birds	Animal host	134	8%	49%	22%	20%	(Gordon and Cowling, 2003)
Mammals	Animal host	497	16%	33%	35%	17%	
Fish	Animal host	12	0%	92%	8%	0%	
Frogs	Animal host	13	8%	85%	0%	8%	
Turtles	Animal host	4	25%	50%	25%	0%	
Snakes and lizards	Animal host	33	15%	70%	6%	9%	
Crocodiles	Animal host	10	20%	70%	10%	0%	
Cow	Animal host	50	28%	58%	0%	14%	(Carlos et al., 2010)
Chicken	Animal host	13	77%	23%	0%	0%	
Pig	Animal host	39	54%	23%	5%	18%	
Sheep	Animal host	29	14%	69%	0%	17%	
Goat	Animal host	16	13%	81%	0%	0%	
Field soil	Nonhost	353	2%	41%	16%	26%	(Bergholz et al., 2011)
Freshwater	Nonhost	190	23%	56%	6%	15%	(Walk et al., 2007)

Interestingly, the great majority of extraintestinal pathogenic strains of *E. coli* (ExPEC) are from phylogroup B2 (Picard et al., 1999; Bingen-Bidois et al., 2002). Epidemiological studies showed that strains with known ExPEC virulence factors are also clustered in phylogroups B2 and to a lesser extent D (Zhang et al., 2002; Johnson et al., 2006). Interestingly, most of the B2-specific virulence factors (iron metabolism, adhesion, lipopolysaccharide biosynthesis) are also involved in the host colonisation process by commensal strains, which led to the idea that pathogenicity was an evolutionary by-product of commensalism in phylogroup B2 (Le Gall et al., 2007). These points strongly suggest that the major ecological strategy of phylogroup B2 is probably evolving towards strict host specialisation (Clermont et al., 2008) and adaptation to the gastrointestinal niche, as compared with other phylogroups. Conceivably, the presence of host-specialised ExPEC in high proportion in this group (if not a sampling bias effect) may contribute to this specialisation via an increased genetic flow within B2 strains.

Conversely, the very low proportion of strains from phylogroup B1 in humans is not observed in wild animals (Gordon and Cowling, 2003) and in nonhost environments such as soil and fresh water (**Table 1.1**) (Walk et al., 2007; Bergholz et al., 2011), which leads to the suggestion that B2 and D strains are host "specialists" (based on their preferred association with humans) whereas A and B1 strains could be considered as host "generalists" (based on their preferred association with non-human hosts and nonhost environments) (Gordon and Cowling, 2003). Although the impact of sampling bias towards samples of human origin seems high in all these reported studies, this trend of host specificity or generalism appears relatively valid when different sampling sources are examined, with the exception of phylogroup A being a

more likely “specialist” than “generalist” (White et al., 2011). Indeed, in a comparative study, strains from both B2 and A were found to harbour key phenotypical traits associated with host-specific life, such as the decreased production of an extracellular matrix and low expression of the major stress regulator protein RpoS (σ^S or σ^{70}) (White et al., 2011).

Phylogroup A seems to be peculiar in its adaptation to the primary environment. It has been observed that strains from group A had the smallest genomes among *E. coli* (Bergthorsson and Ochman, 1998). Incidentally, fewer accessory genes are found in phylogroup A, in which core genes are also found in all other phylogroups (Sims and Kim, 2011). This phenomenon has been characterised as “commensal minimalism”, and of which *E. coli* K-12 strains (belonging to phylogroup A) are a perfect example: increased levels of host-adaptation usually leads to increased shedding of unnecessary pathogenesis-related genes (Moran, 2002). The apparent effect of commensal minimalism is a genome size reduction, the extreme of which is observed in endosymbionts or obligate intracellular organisms. Generally speaking, bacteria with the biggest genome size are ecologically associated with the necessity to adapt to short-term changing of their living conditions, requiring a certain plasticity that is achieved by maintaining a large number of diverse genes (Ochman and Davalos, 2006). Opportunistic pathogens (such as most of *E. coli* pathogens) have usually a genome comprising between 2 and 5 Mb, which is considered as an average among bacteria (Ochman and Davalos, 2006). The best illustration of this commensal minimalism hypothesis is that very few pathogenic strains are found in *E. coli* phylogroup A, which is not the case for any other phylogroup (Sims and Kim, 2011). It has then been suggested that, similarly to obligate pathogens, phylogroup A may be

evolving towards “obligate commensalism” (Sims and Kim, 2011). To contrast this hypothesis, it has very recently been shown that a strain from phylogroup A was able to cause mastitis in bovines (Dufour et al., 2011), but more genomic characterisation needs to be done on more samples to determine how these pathogenic strains fit with the “commensal minimalism” model.

As we hope to have shown in this section, the analysis of genome dynamics and its link with bacterial population structure is crucial to get an integrated view of *E. coli* ecology. It is equally important to understand what are the major molecular and physiological mechanisms involved in environmental adaptation.

1.2. The ecology of environmental adaptation in *E. coli*

Two major dichotomies can be drawn when addressing the topic of environmental adaptation for *E. coli*, reflecting different lifestyles this bacterium can adopt. In this section, we give an overview on the differences and similarities between pathogenic and commensal lifestyles in *E. coli*, and also on the host (primary environment) or nonhost (secondary environments) mechanisms of association, and possibly adaptation.

1.2.1. Pathogenic and commensal *E. coli*

1.2.1.1. Functional and genetic diversity of E. coli *pathogens*

The primary environment of *E. coli* is the gastrointestinal tract of warm-blooded animals, that it colonises only hours after birth, but it can also be excreted *ex vivo* via faecal matter in secondary environments as diverse as the ecology of the primary hosts allows. Around this basic ecological framework, multiple adaptive strategies are observed in *E. coli*, all linked to survival and transmissibility abilities. Most of the time, *E. coli* persists asymptomatically in its hosts, without any obvious effect on their health or physiology. However, a surprising diversity of pathogenic strains (or "pathotypes", or "pathovars") has been described in the last 30 years as pathogenic *E. coli* can cause various types of intestinal or extra-intestinal infections in humans (**Figure 1.3**) via a very diverse set of virulence factors, affecting a range of cellular processes (Kaper et al., 2004). These virulence factors are acquired laterally, via the transfer of pathogenicity islands or recombination and are mostly involved in the

colonisation process, providing fitness-increasing metabolic abilities, motility, adhesion and persistence mechanisms (Kaper et al., 2004). The diversity of the different pathologies makes *E. coli* an important pathogen worldwide, with an estimated burden of around 2 million child deaths per year caused mainly by diarrhoeal infections and sepsis after urinary tract infections (Touchon et al., 2009) as well as a considerable economic impact (Russo and Johnson, 2003).

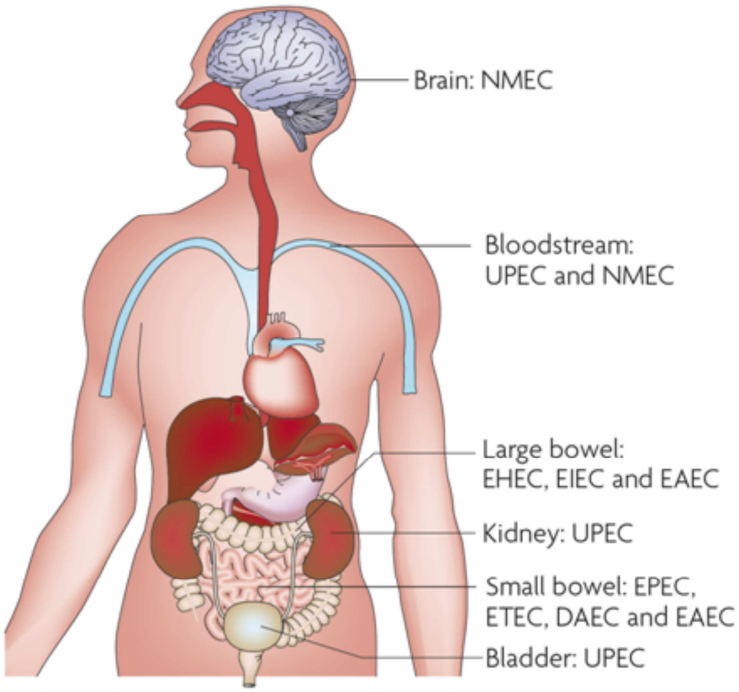


Figure 1.3. Human body sites of colonisation by various *E. coli* pathotypes (Croxen and Finlay, 2010)**.** "EC" in acronyms refers to "*Escherichia coli*" with the different pathotypes being, for intestinal pathogens: "enterohaemorrhagic" (EHEC), "enteroinvasive" (EIEC), "enteroaggregative" (EAEC), diffusely adherent" (DAEC); and for extraintestinal pathogens (ExPEC): "neonatal meningitis" NMEC, "uropathogenic" (UPEC). This figure is copyrighted by MacMillan Publishers Ltd.

Intestinal pathologies caused by *E. coli* are generally observed after infection by well-described pathotypes. The best-studied *E. coli* intestinal pathogenic strains, popularised by their association with meat and vegetables, are *E. coli* from serotype

O157:H7, which is the archetype of the enterohaemorrhagic *E. coli* (EHEC) pathotype.

Strains of the O157:H7 serotype are predominant EHEC pathogens in North America, UK and Japan, but several other serotypes, particularly those of the O26 and O111 serogroups, can also cause disease and are more prominent than O157:H7 in many countries (Kaper et al., 2004). The ability of producing two families of Shiga-toxins, Stx1 and Stx2 that can spread systemically and, upon entry of host cells, disrupt protein synthesis by cleaving the 28S ribosomal subunit (Donohue-Rolfe et al., 1991), and the presence of a group of genes on a 35-kb pathogenicity island called the locus of enterocyte effacement (LEE) are the two factors required for human toxicity of *E. coli* O157:H7 strains (O'Brien and Holmes, 1987; McDaniel et al., 1995). Other non-O157 strains can produce Shiga or Shiga-like toxins and are usually regrouped under the appellation of Shiga toxin-producing *E. coli* (STEC) or verotoxin-producing *E. coli* (VTEC). In 10% of the cases, mostly in immuno-compromised patients, children and the elderly, the intestinal diarrheic infection by STEC can evolve into haemolytic-uraemic syndrome (HUS), leading to acute renal failure, haemolytic anaemia and possible death. Most EHEC strains contain the LEE, which encodes transcriptional regulators, the adhesin intimin, a quite conserved filamentous type III secretion system (T3SS), chaperones, translocators (EspA, EspD, EspB) and six effector proteins (Garmendia et al., 2005). Other intestinal pathotypes of *E. coli* have different mechanisms of infection, ranging from the production of other enterotoxins [enteropathogenic *E. coli* (EPEC) and enterotoxigenic *E. coli* (ETEC)], increased adhesion and biofilm formation on the epithelial layer [enteroaggregative *E. coli* (EAEC), diffusely adherent *E. coli* (DAEC)]. More details on the molecular

mechanisms of infection by these pathogens have been presented in very good reviews (Kaper et al., 2004; Croxen and Finlay, 2010).

The two most important extraintestinal pathogenic *E. coli* (ExPEC) pathotypes [neonatal meningitis *E. coli* (NMEC) and uropathogenic *E. coli* (UPEC)] are able to colonise the bloodstream (UPEC, NMEC), bladders and kidneys (UPEC) and the brain (NMEC) (**Figure 1.3**). The process of infection by NMEC is complex as the bacteria need to transfer and live in the bloodstream, actively cross the blood-brain barrier and persist in the cerebrospinal fluid (Croxen and Finlay, 2010). UPEC are able to live in the bloodstream (Smith et al., 2010) from which they can infect kidneys and more frequently the bladder, being the most common cause of urinary tract infections (UTI). Other ExPEC include *E. coli* veterinary isolates causing mastitis in cattle by persisting in udders without being killed by the immune system (Buitenhuis et al., 2011), and for which specific genomic islands have been recently observed (Dufour et al., 2011). Much rarer ExPEC-related pathogenicity mechanisms have recently been described, such as the production of "flesh-eating" dermonecrotic toxins (Grimaldi et al., 2010), or the infection of the endometrium (i.e., the inner membrane of the mammalian uterus) by a newly-described endometrial pathogenic *E. coli* (EnPEC) (Sheldon et al., 2010). However, more research on these strains is required to understand fully how they can cause disease, and what their ecology is.

E. coli pathogens share evolutionary similarities and are believed to evolve mainly via two non-exclusive mechanisms: (a) the acquisition via HGT of genes or pathogenicity islands containing fitness-enhancing virulence factors and sometimes (b) the apparition of deletions, or "black holes" (Maurelli et al., 1998; Dobrindt and Hacker,

2001). Conceivably, these two mechanisms are occurring on different time scales. Environmental fitness can change drastically and suddenly upon the acquisition of pathogenicity islands but, it could take longer evolutionary time for “black holes” to be fixed and selected for. Similarly to the “obligate commensalism” hypothesis presented in the previous section, it is commonly accepted that *Shigella* and EIEC are evolving toward a more intimate, obligate pathogenesis, for which genome reduction is positively selected as the loss of functions improves the overall fitness of the organism (Maurelli et al., 1998). As a matter of fact, *Shigella* and EIEC are the only obligate pathogens among *E. coli* pathogens, which are otherwise considered as “opportunistic” as they can persist for a long time in healthy hosts (Touchon et al., 2009).

The acquisition via HGT of virulence genes harbouring mobile genetic elements, such as genomic islands or plasmids, seems to be a common feature in many *E. coli* pathotypes (**Figure 1.4**). For instance, all EIEC and *Shigella* strains harbour a 230-kb pINV plasmid that is strictly required for their pathogenicity (Lan et al., 2003). EPEC and ETEC virulence is greatly dependent on the presence of pEAF and pENT plasmids, both harbouring toxin genes (**Figure 1.4**). UPEC strains possess an increased number of iron acquisition systems located on genomic islands, a signature of their adaptation to the urinary tract (Lloyd et al., 2007). Virulence factors of *E. coli* O157:H7 are mostly encoded by chromosomally inserted genomic islands (e.g., LEE), prophage elements (Shiga toxins) or plasmids (various toxins) (Law, 2000).

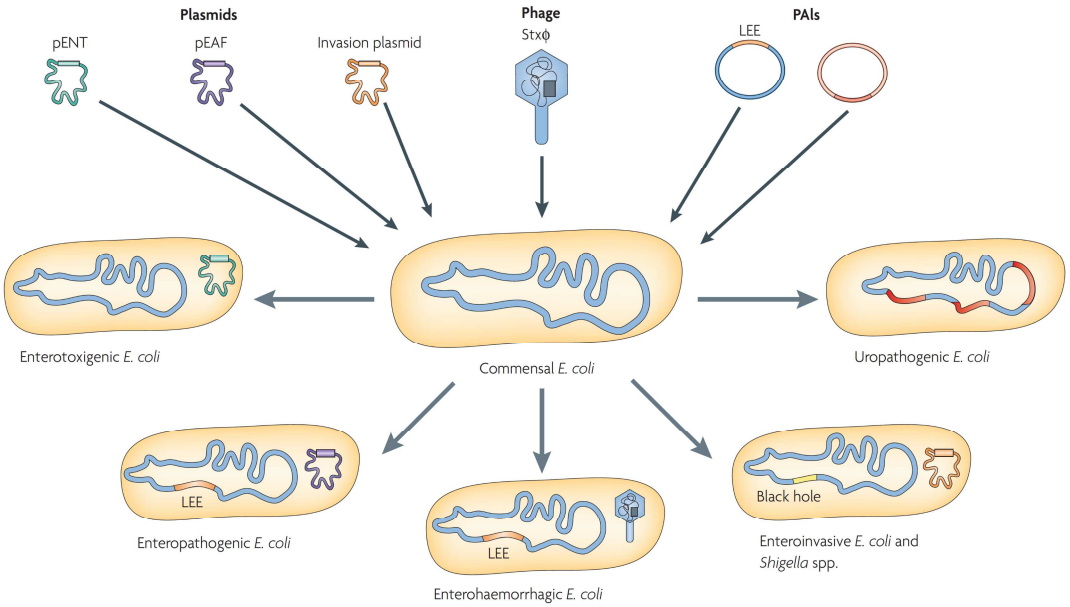


Figure 1.4. Contribution of HGT to the evolution of different *E. coli* pathotypes (Ahmed et al., 2008). pENT and pEAF are the plasmids conferring virulence to ETEC and EPEC, respectively. The invasion plasmid (pINV) confers invasive properties to EIEC and *Shigella*. LEE: locus of enterocyte effacement; PAIs, pathogenicity-associated islands; StxΦ: bacteriophage harbouring Shiga toxin genes. See text for more information. This figure is copyrighted by the Nature Publishing Group.

Upon acquisition of genomic islands harbouring virulence factors, a commensal *E. coli* bacterium can in theory become pathogenic and change drastically its lifestyle. It has been suggested that the acquisition of virulence was not identical for all *E. coli* leading to the idea that certain phylogroups, such as B2 and D, were more associated with ExPEC pathogenicity than others (Escobar-Paramo et al., 2004). Similarly, the presence of similar toxins in closely related *Shigella dysenteriae* serotype 1 and *E. coli* O157:H7 seems to indicate that a specific genetic background is required to express those toxins (Touchon et al., 2009). This view is contrasted by the fact that various pathogens are found to be from almost all phylogroups (Sims and Kim, 2011). As the pathogenesis develops, a successful pathogen has to escape host immunity and defences. Because of this, pathogenic *E. coli* presumably have drastically different

ecological strategies compared to commensals, as they must cope with additional environmental pressures linked to their pathogenic lifestyle.

Immune evasion seems to be one of the major evolutionary constraints on pathogens (Frank and Schmid-Hempel, 2008), as reflected by the observed variety of mechanisms to avoid host defences. For instance, 80% NMEC strains possess a K1 capsule providing them with a physical protection against lysozymal fusion and therefore aids during the crossing of the blood-brain barrier (Croxen and Finlay, 2010). NMEC strains also harbour a prophage-encoded acetyltransferase that modifies the O-antigen as a way to escape immune recognition (Deszo et al., 2005). Additionally, among all sorts of *E. coli* pathotypes, a huge variety of adhesins, flagellar proteins and other fimbriae are observed to be involved in the pathogenesis process itself (Le Bouguenec, 2005), highlighting the importance of motility during host colonisation by pathogens, something which is not necessarily obvious for commensals. Interestingly, it appears that the processes involved in nonhost persistence and host transmission are not entirely specific to pathogens (van Elsas et al., 2011), suggesting that the evolutionary constraints for transmissibility and nonhost adaptation are shared between pathogens and commensals. This point is supported by the fact that many “virulence-associated” functions have been found to also be important for intestinal persistence in commensals (Wold et al., 1992; Lipsitch and Moxon, 1997) indicating that virulence is probably not the major selective force on these functions.

1.2.1.2. Common themes between pathogenic and commensal **E. coli**

Examples of the functional overlap between traits associated with virulence in pathogens and colonisation in commensals have been identified when comparing multiple genomes (Rasko et al., 2008; Touchon et al., 2009). For instance, the examination of the genome of *E. coli* strain HS, an archetypical commensal strain with few assumed laboratory adaptation and high host colonisation abilities, has shown that it shares with pathogenic strains many genes previously thought to be virulence factors, notably pilus and fimbriae genes involved in colonisation (Rasko et al., 2008). Interestingly, the presence of a type II secretion system in *E. coli* HS, with no apparent link to mobile genetic elements suggests that it could represent a true niche specialisation adaptation and not a "random" gene transfer (Rasko et al., 2008). A similar comment can be drawn from the presence of ETT2 in strain HS, a non-functional type III secretion system used as a specific marker for *E. coli* pathogens (Rasko et al., 2008).

As pathogens are understandably more heavily scrutinised by microbiological research, many laterally-acquired genes such as those harboured by genomic islands are thought to be important for virulence, as knockouts impair colonising fitness and thus the progression of infection. However, the detection of these islands in strains with no obvious pathogenetic behaviour underlines the multiple functional roles of genomic islands. Indeed, many genomic islands can be functionally characterised, for example as ecological, saprophytic, symbiotic or pathogenicity islands (Hacker and Carniel, 2001). This is for instance the case with the genes encoding the production of yersiniabactin, an iron-scavenging siderophore molecule. It is located on a genomic

island that was observed to be an important colonisation factor for UPEC in the urinary tract. Genomic analyses showed that yersiniabactin is present in numerous non-pathogenic *E. coli* strains, with nevertheless a higher presence in phylogroups B2 and D, suggesting an ecological relevance in host-specialisation and perhaps the early acquisition by a B2 ancestor (van Elsas et al., 2011). Interestingly, another very good host coloniser, the probiotic strain *E. coli* Nissle 1917 (Altenhoefer et al., 2004; Ukena et al., 2007; Schultz, 2008) is able to produce the 4 different siderophores described for *E. coli* (Valdebenito et al., 2006) including yersiniabactin, suggesting that this trait is involved in host adaptation by *E. coli* in general and not just in pathogens.

The mosaic structure of *E. coli* commensal genomes shows that most of the virulence factors are in fact colonisation functions and that it is the horizontal genetic flow producing the right combination of these factors that makes a successful pathogen. Interestingly, this flow is not unidirectional: commensals can become pathogens, but the contrary is also observed. *E. coli* strain 83972 was isolated from the urinary tract of a young girl who had carried it for 3 years without symptoms (Hancock et al., 2008). It was observed that 83972 was a very good coloniser of the urinary tract, and could even outcompete some UPEC in urine (Roos et al., 2006). Subsequently, this phenomenon was found to be frequent and it was observed that these "asymptomatic bacteriuria" (ABU) strains were phylogenetically related to UPEC but had experienced genome reduction events (Zdziarski et al., 2008), which probably reflects their ongoing host specialisation (as previously presented in this chapter). The phenomenon of "commensal conversion" captured by the examination of strain 83972 and other ABU strains is likely to be widespread even among intestinal *E. coli*, as we

have an inevitable sampling bias towards successful pathogens (how can one sample specifically for commensal strains that were once pathogenic?). The fact that we observe a mosaic of virulence-associated or colonisation-associated genes in *E. coli* genomes is an indirect indication that there could be a constant balance across evolutionary times between pathogenic and commensal status in *E. coli* strains.

This point is supported by the observation that when phylogenies are reconstructed, pathogens and commensals are generally found to belong to the same phylogenetic backgrounds, with no obvious distinction between strains with different lifestyles. This observation suggests that in most cases, pathogenicity is not a strong speciating force, with functions associated with all kinds of *E. coli* being positively selected for. In other words, from an evolutionary and global point of view, the difference between pathogens and commensals is very small as within the open pangenome of *E. coli*, very few traits seem to be pathogen-specific. It seems indeed logical that most of the time (and probably never for the majority of them), facultative *E. coli* pathogens do not express pathogenicity-specific genes but are however committed to the same requirements as commensal *E. coli* for survival, replication and general physiology. Notably, this point may be different for EIEC and *Shigella*. Accordingly, it has been observed that “black holes” (i.e., large deletions observed consistently in *Shigella* and EIEC genomes) are constituted by the deletion of important functions such as amino acid and carbohydrate transport or nucleoside metabolism (Touchon et al., 2009). These functions are believed to be important in the metabolic plasticity and environmental adaptation flexibility of *E. coli*, strongly indicating that *Shigella* and EIEC are following an obligate host-association evolutionary pathway, moreover at a seemingly accelerated rate (van Passel et al., 2008).

The fact that pathogenic and commensal lifestyles have coevolved simultaneously within the same species, probably since the divergence of the *E. coli* ancestral progenitor, implies that this duality was probably under positive selection as it gave a presumed ecological advantage over other competing intestinal bacteria, most of which do not seem to have pathogenic variants. In this part, the ecological dichotomy between these two lifestyles was briefly detailed, with an emphasis on the possible impact they had on *E. coli* evolution. There is another major ecological dichotomy in *E. coli*, introduced almost 30 years ago (Savageau, 1983). It is crudely estimated that because of the inevitable faecal shedding of *E. coli*, half of all existing strains are associated with secondary non-intestinal, nonhost environments. The potential impacts on *E. coli* ecology and evolution are presented in the next section.

1.2.2. The ecology of *E. coli* host and nonhost environmental persistence

1.2.2.1. Life in the primary ("host") gastrointestinal tract environment

The primary niche of *E. coli* is the mucus layer of the lower intestine of mammals, comprising the colon and distal parts of the ileum. It has been reported that when embedded in the intestinal mucus layer, *E. coli* had a generation time of 40 to 80 min (Poulsen et al., 1994; Poulsen et al., 1995). There are 100 to 1,000 times more *E. coli* in the colon than in the ileum, yet some arguments based on lactose utilisation and intestinal flow models have been advanced to suggest that *E. coli* was more adapted to life in the small than large intestines, and that the observed high concentrations in

the colon were due to mechanical displacement during digestion rather than active adaptation processes (Koch, 1987). Increased colonisation of the ileum by some strains of *E. coli* has also been observed (Staley et al., 1969; Barnich et al., 2007).

Nevertheless, most ingested *E. coli* strains are thought to transit through the mammalian gut without much effect on their host (Caugant et al., 1981; Savageau, 1983). Some strains are able to persist for a few months or years, while others are readily and constantly excreted to the external nonhost environment. This observed duality between "residents" and "transients" has been made in early host-association studies (Wallick and Stuart, 1943; Sears et al., 1950; Sears and Brownlee, 1952; Cooke et al., 1972; Smith, 1975) using serotyping as a method of discrimination between strains. More recently, and as mentioned in section 1.2.1 of this thesis, a study distinguished resident and transient strains in infants according to their phylogenetic group and it was observed that phylogroup B2 strains were more likely to be resident whereas other phylogroups could be more considered as transient (Nowrouzian et al., 2005). In contrast to mucus-embedded cells, *E. coli* living in the luminal contents have not been observed to grow, and are believed to be excreted with faeces (Poulsen et al., 1995). This differential ability in terms of mucus colonisation, whether active or passive, may be the basis of the dichotomy between "resident" and "transient" strains.

The main reason why active viable bacteria persist in one given environment is because they can scavenge and use suitable nutrients for the maintenance of their physiology. Therefore, metabolic abilities are important to consider when examining environmental adaptation, especially in the gut where the global dominant gene

expression by resident bacterial communities is in majority linked to carbohydrate metabolism (Booijink et al., 2010). *E. coli* K-12 has been shown to use micro-aerobically and anaerobically specific C-sources in mucus-based medium mimicking conditions in the intestine (Chang et al., 2004; Jones et al., 2007). Gluconate was the preferred source of carbon, followed by *N*-acetylglucosamine, *N*-acetylneuraminic acid and glucuronate, mannose, fucose and ribose (Chang et al., 2004; Alpert et al., 2009). Fucose and ribose share common metabolic processes and were used dynamically by *E. coli* in the gut (Autieri et al., 2007). Glycogen could potentially also be used, since mutants unable to synthesise or store glycogen have reduced colonisation abilities (Jones et al., 2008). Apparently, not only catabolic functions are important, as mutant strains impaired in purine and pyrimidine biosynthesis were eliminated from the mouse intestine (Vogel-Scheel et al., 2010). Additionally, maybe not all metabolic functions important for intestinal colonisation are commensal functions in *E. coli*. It was shown that EHEC and *E. coli* K-12, albeit using similar carbon sources in the mouse gut, were also using specific ones (Fabich et al., 2008). Additionally, EHEC strains seem to gain advantage from using novel metabolic compounds, such as ethanolamine as a nitrogen source in the cattle intestines (Bertin et al., 2011). Interestingly, *E. coli* O157:H7 colonised a sterile mouse intestine successfully but was outcompeted in mice pre-treated with *E. coli* K-12 (Miranda et al., 2004). *E. coli* O157:H7 showed less metabolic flexibility *in vivo* than *E. coli* K-12, which seemed, in this study, more metabolically adapted to host conditions (Miranda et al., 2004).

Other mechanisms have been pinpointed as important during the colonisation by *E. coli* of primary host environments. Type-1 pili are important for the initial attachment

to the intestinal surface and the colonisation of mucosa (Tullus et al., 1992; Herias et al., 1995), and are harboured by commensal and UPEC strains (Hartl and Dykhuizen, 1984). The presence of an extracellular capsule (antigen K5) and of P fimbriae also increases intestinal colonisation by *E. coli* in gnotobiotic rats (Herias et al., 1997). Flagellar motility seems however to be impairing colonisation, with nonmotile O157:H– EHEC variants colonising cattle intestines very much better than motile ones (Dobbin et al., 2006). This observation was confirmed by a study using "intestine-adapted" strains derived from *E. coli* K-12 MG1655 which were observed to become nonmotile only a few days after feeding mice (Gauger et al., 2007). This lack of motility upon host colonisation was associated with mutations in the *flhDC* regulator (Dobbin et al., 2006; Gauger et al., 2007; De Paepe et al., 2011).

1.2.2.2. Secondary ("nonhost") environments

Interestingly, although attachment through pili and fimbriae seems important, biofilm formation (i.e., the formation of an extracellular adhesive matrix) is not considered to be an advantage during intestinal colonisation. The production of the matrix components is very tightly regulated in *E. coli* and *Salmonella* and responds to many environmental signals (Barnhart et al., 2006). Notably, matrix expression is generally high at low temperatures (28-30°C) and inhibited at host temperature (37°C) (Arnqvist et al., 1992; Barnhart et al., 2006). Several other conditions preferably linked to ex-vivo conditions have been identified to enhance matrix production (Barnhart et al., 2006). Accordingly, it has been shown that the formation of an extracellular matrix in *Salmonella* (which is very similar to the matrix of *E. coli*) was not favoured during intestinal colonisation (White et al., 2008). Instead, matrix

formation conferred a higher resistance to desiccation (Gibson et al., 2006) and long-term survival (White et al., 2006), clearly highlighting its role in inter-host transmission or environmental survival rather than host colonisation.

Much sampling effort in pathogenic and non-pathogenic *E. coli* ecology studies has been concentrated on animal and human hosts only. However, it has been suggested that half of all living *E. coli* could theoretically live outside hosts, in secondary environments such as soil, water and sediments (Savageau, 1983). In his landmark 1983 review, Michael Savageau presents different hypotheses on host and nonhost lifestyle adaptations in *E. coli* (Savageau, 1983). *E. coli* originates from the primary environment, an intestine, and is excreted via faecal matter in environments congruent with the warm-blooded animal host ecology, most of the time water, sediments or soils. A large majority of strains probably die in this step, and a fraction of the survivors will presumably be able to recolonise an intestine (Savageau, 1983; Winfield and Groisman, 2003). However, some strains are found to persist better (or in higher proportions) in secondary environments than others (we call them “nonhost-associated” strains), which indicate that life in these environments is not a random, accidental event but probably has an ecological significance on the species as a whole. Savageau hypothesised that species-wide selective pressures in *E. coli* were maintaining a high growth rate in hosts and a long half-life in nonhost environments (Savageau, 1983). This view is very consistent with the host specialisation and generalisation mechanisms introduced in section 1.1.3.2.

If *E. coli* is able to live in ecologically important host and nonhost environments, it must be reflected in a variable association with specific traits. The acquisition of

nutrients is a good candidate, as different nutrients are presumably available in primary and secondary environments (Savageau, 1983; Winfield and Groisman, 2003). It is also possible that the abundance of different nutrients vary, with secondary environments harbouring less available compounds (Savageau, 1983). Accordingly, different ecological processes are thought to influence the requirements for transition between the primary and secondary environments for *E. coli*. The "demand theory" (Savageau, 1974; Savageau, 1983) implies that environmental changes can influence how core metabolic genes are regulated, balancing between positive regulation for high-demand gene products and negative regulation for low-demand gene products in a way that could mirror energy source availability in natural environments. Alternatively, the "selection theory" suggests that some isolates are fitter than others (i.e., harbour additional traits) for life in particular host or nonhost environments and thus retrieved in higher quantities when these environments are sampled (Whittam, 1989; Gordon et al., 2002). These proposed ecological processes are neither exhaustive nor exclusive; it is likely that a mixture of core gene regulation variation and presence or absence of fitness-enhancing traits governs environmental adaptation.

Interestingly, the differential abilities among *E. coli* strains to acquire nutrients has also been linked with mutations affecting the expression levels of the alternative sigma factor RpoS (σ^S or σ^{38}), which can be found in a relatively large proportion of natural *E. coli* isolates (Waterman and Small, 1996; Bhagwat et al., 2005; Ferenci, 2005; Ferenci et al., 2011). Strains with low RpoS levels were found to be more able to compete for nutrients with other bacteria, but were less capable of surviving stresses such as acid shock or starvation (King et al., 2004). This led to the hypothesis

that *E. coli* faces a trade-off between self-preservation and nutritional competence (SPANC) and suggested that strains with different positions on the SPANC balance might occupy different ecological niches (Ferenci, 2005) or experience selective pressures to maintain diversity (Levert et al., 2010; De Paepe et al., 2011). It is conceivable that the balance between host and nonhost environments parallels the SPANC balance, as it has been shown that this mechanism could promote strain diversification in the mouse gut (De Paepe et al., 2011) with possibly different ecological behaviours.

1.3. *E. coli* lifestyle in agricultural fields

1.3.1. Persistence and agricultural sources of contamination

1.3.1.1. Increased persistence of E. coli *in soil and water in relation to population structure*

As *E. coli* is being shed actively and in tremendous amounts from animal hosts, at least detectable concentrations of *E. coli* should be found in natural environment, especially in agricultural areas. This point is the basis of the use of *E. coli* as a faecal indicator in water, soils and the food industry. Faecal indication relies on the fact that *E. coli* precisely has a high rate of die-off and does not persist long enough in secondary environments so that it can accurately predict faecal contamination. The validity of this requirement has increasingly been challenged by the report that *E. coli* could be isolated from secondary environments independently of seasonality or obvious source of faecal input. For instance, long-term persisting *E. coli* strains have been isolated from undisturbed forest soils (Byappanahalli et al., 2006). When DNA fingerprint profiles between strains from this forest soil and the surrounding wildlife were compared (and thus the genomic rearrangements patterns between strains as explained in section 1.1.2.), soil strains formed a distinct cohesive group, whereas strains from wild animals did not cluster in any way (Byappanahalli et al., 2006). This suggests that the observed *E. coli* forest soil populations are possibly autochthonous and might form a distinct persisting population in soil. Numerous other studies in different geographical locations have reported the same observations of an autochthonous, non-transient presence of so-called “naturalised” *E. coli* persisting and

even growing in tropical soils (Byappanahalli et al., 2006; Ishii et al., 2006; Ishii and Sadowsky, 2008; Goto and Yan, 2011), water (Bermudez and Hazen, 1988; Power et al., 2005; Vital et al., 2008) and sediments (Solo-Gabriele et al., 2000; Whitman and Nevers, 2003; Whitman et al., 2003; Ishii et al., 2007) at low but still detectable levels. These "naturalised" populations of *E. coli* probably have to be considered when interpreting faecal indication tests in tropical environments (Goto and Yan, 2011).

It has been suggested that the increased number of studies reporting "naturalised" *E. coli* in tropical environments were probably reflecting the fact that tropical environmental conditions "mimicked the colon environment" (Winfield and Groisman, 2003). However, this reductionist approach is probably too simple to be true, as intestinal contents are tremendously more different in microbial phylogenetic composition, density and physicochemical constraints than an open environment such as water or soil, in any type of climate. Moreover, recent reports describe the presence of environmentally persistent *E. coli* in low-temperature Irish soils, which are far from being tropical (Brennan et al., 2010; Brennan et al., 2010). "Naturalised" *E. coli* were also found to grow in watershed soils and beach sands from the temperate Lake Superior region in USA (Ishii et al., 2006; Ishii et al., 2007; Ishii et al., 2010). Another contradicting and surprising example comes from alpine pasture soils in the French Alps, in which *E. coli* was expectedly detected during the cattle grazing season but more surprisingly all-year long, even when a snow layer had formed, and thawed (Texier et al., 2008). Finally, in the most longitudinal study ever performed to our knowledge, *E. coli* was found to persist at low levels for about 13 years (from 1978 to 1991) in experimentally inoculated rye-grass soils in Vermont, USA

(Sjogren, 1995). These studies indicate that far from being restricted to tropical environments, it is likely that faecal deposition events can generate self-replicating, sustainable, low-density populations of soil-borne *E. coli* that can be considered as part of the resident soil microbiota (Byappanahalli et al., 2006; Ishii et al., 2006).

In any case, why does the majority of "naturalised" *E. coli* seem to be associated to soils? First, this could be a bias linked to increased sampling in soils, but there could also be a real ecological reason. *E. coli* is mostly present in the intestines of birds and endothermic wildlife, which live predominantly on land. Thus soils and to a lesser extent freshwaters, are the most likely environments to be contaminated by faecal matter and drive adaptation, if any, of nonhost-associated or "naturalised" strains. In that sense, vegetation can also be considered as important, yet no population-wide study has ever been reported to our knowledge, placing this thesis work in a good context of novelty. Another reason, linked to the previous one, explaining why "naturalised" *E. coli* would be maintained predominantly in soils would be that environmental conditions are somehow suitable for *E. coli* in this milieu. Indeed, some strains of *E. coli* possess very complicated machinery for the biodegradation of aromatic compounds, with up to 5 distinct catabolic pathways (Diaz et al., 2001). Aromatic compounds are supposed to be an obvious nutrient for *E. coli* colonising soils, water and plants, given their high concentration in those environments (Diaz et al., 2001). Aromatic compounds can also be present in the gut, in the form or aromatic amino acids, such as tryptophan, and sometimes steroids and drugs (Diaz et al., 2001) but as presented before, the preferred sources of nutrients in the intestines for *E. coli* do not seem to include aromatic compounds (Chang et al., 2004; Alpert et al., 2009), which may then only be of minor ecological importance in the gut. In that respect, it

would be interesting to examine how the aromatic compound metabolic abilities are distributed in nonhost-associated strains, and whether it could be considered as an important ecological factor. This point is examined further in Chapter 4 of this thesis.

Unfortunately, most of the studies assessing these large populations of "naturalised" isolates did not address the question of population structure. In the future, this interesting information will shed more light on the global ecological processes occurring in *E. coli*, and determine if, similarly to a "host specialisation" process, there could be an "increased generalisation" process in *E. coli*. Additionally, it would be interesting to examine if the acquisition of specific traits could be associated with this long-term persistence phenotype. A few studies examine more in depth the population structure of strains isolated in freshwater and pasture soil (Walk et al., 2007; Bergholz et al., 2011). There is no indication on the "naturalised" status of the tested strains but it is nevertheless interesting to examine whether the population structure of persisting nonhost-associated *E. coli* can be linked to their ecology. Interestingly, the majority of *E. coli* isolates in these two studies are from phylogroup B1 (41% in pasture soil (n=353); 56% in freshwater (n=190); see **Table 1.1**). Few studies on the population structure of nonhost-associated *E. coli* are available, but this strong consistency in B1 dominance prompts the eventuality that phylogroup B1 strains may harbour traits that grants them a competitive advantage or an increased survival in secondary environments (Walk et al., 2007). We discuss this topic further along with our research in Chapter 4 of this thesis.

1.3.1.2. **E. coli** *from wildlife and domesticated animals*

An alternative hypothesis explaining the presence of certain genotypes or phylogroups in nonhost environments could simply be their presence in the dominant contaminating sources. This would imply that nonhost environments have little effect on the *E. coli* population structure. As nonhost environments are so different in terms of physicochemical conditions, this seem unlikely but accordingly, it has been observed that wild birds (Gordon and Cowling, 2003) and farm animals (Walk et al., 2007; Carlos et al., 2010) were generally populated by a large majority of strains from phylogroup B1 (**Table 1.1**). Coupled with the observation that soil and water also harboured more strains from phylogroup B1, the above hypothesis about contaminating sources could be true. Nonetheless, more longitudinal studies focusing on the source of isolation are required. Interestingly, it has recently been suggested that phylogroup B1 was composed of "host generalists", with no specific host association preferences as opposed to phylogroup B2, composed of "host specialists" (White et al., 2011). It could be that host generalism in *E. coli* is associated with nonhost adaptation. We also address this topic in Chapter 4 of this thesis.

1.3.1.3. Possible sources and vectors of agricultural field contamination

As briefly mentioned in the previous section, the most heavily contaminated natural environments by faecal input from most endothermic species are presumably soils, freshwaters and plants, making *E. coli* contamination in agricultural fields inevitable. Indeed, low levels of *E. coli* are constantly observed when analysing vegetable

samples during faecal indication tests (Ibenyassine et al., 2007; Rai and Tripathi, 2007; Ilic et al., 2008; Valentin-Bon et al., 2008; Mandrell, 2009; Caponigro et al., 2010; Oliveira et al., 2010), suggesting that the association with plants in particular is not uncommon. Anthropic or natural contamination of agricultural fields by *E. coli* can occur in multiple ways, the most obvious sources being the water used for irrigation, and direct (wildlife, humans) or indirect faecal contamination resulting from untreated manure used for fertilisation (**Figure 1.5**).

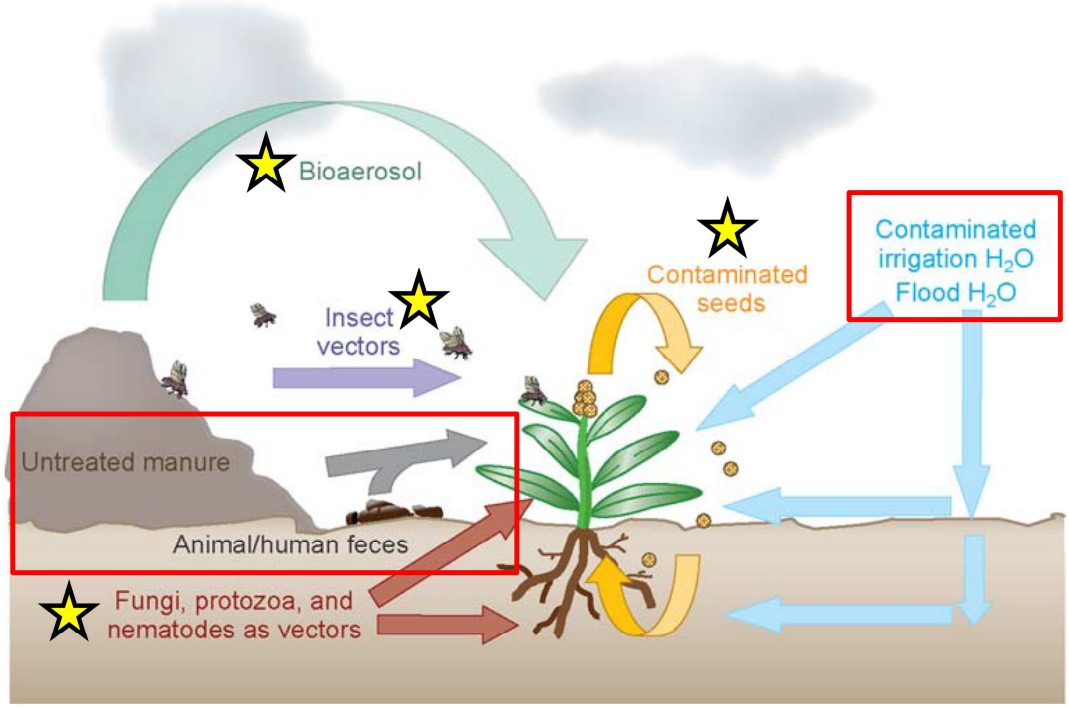


Figure 1.5. Possible routes of agricultural field contamination by *E. coli* (Brandl, 2006). Red squares represent the major sources of *E. coli* contamination and yellow stars represent the factors contributing to the spread (vectors) of *E. coli* from the major sources. This figure is copyrighted by Annual Reviews.

Water is one of the most likely sources of *E. coli* contamination in agricultural fields (Steele and Odumeru, 2004), and a number of outbreaks of pathogenic *E. coli* have been linked to fresh produce contaminated by irrigation (Rice et al., 1992; Soderstrom

et al., 2005). In the UK, 71% of salad fields are irrigated from surface waters (Tyrrel et al., 2006) which likely received treated wastewater effluents. Levels of coliforms in irrigating water are monitored on very regular bases, with an accepted threshold of less than 10^3 cells per 100 ml (Tyrrel et al., 2006), which is nevertheless conducive of a low-level contamination in agricultural fields. *E. coli* was also detected on lettuce leaves 30 days after it had been contaminated by a single application of contaminated water, with a possible growth in the phyllosphere (Solomon et al., 2003). Additionally, *E. coli* could internalise in the inner tissues of spinach leaves when contaminating water, and not soil, was applied (Mitra et al., 2009) indicating that plants contaminated from above may increase the likelihood of *E. coli* resisting surface decontamination procedures, as compared to contamination from the roots. However, this point is contrasted by the observation that roots experimentally inoculated with *E. coli* resulted in the systemic presence of the bacterium endo- and epiphytically (Cooley et al., 2003), suggesting that plants can be colonised by *E. coli* from the soil.

Indeed, field soil is also a potential source of contamination by *E. coli* which, in addition to the potentially "naturalised" resident populations described in section 1.3.1.1, can likely come from the dispersion of animal wastes such as slurry or manure. Typically, and depending on various conditions, the concentration of *E. coli* in faecal matter can naturally vary between 10^2 and 10^5 cells/g, in slurry between 10 and 10^4 cells/g and in manure between 10^2 and 10^7 cells/g. A very large number of studies have focused on the survival abilities of *E. coli* in soils after manure application (Whipps et al., 2008). Results seem to vary according to the experimental

protocol but the overall trend is that persistence can be long, ranging from 10 days to several years (Whipps et al., 2008; Mandrell, 2009).

In addition to these two major sources of contamination, a multitude of possible vectors (or more "minor" sources) have been identified. Insects such as flies (Janisiewicz et al., 1999; Sasaki et al., 2000; Sela et al., 2005) and honeybees (Johnson et al., 1993) have been shown to contribute to the spread of phytopathogens and *E. coli* on plants. Flies may constitute an under-represented source of *E. coli* transmission, as they can travel for relatively long distances and are attracted to faecal matter and farm animals. Additionally, when in natural environments, bacteria have to cope with the constant predation by nematodes and protozoa living abundantly in soils. Subsequently, it has been shown that *Caenorhabditis elegans* could carry live *E. coli* and *Salmonella* in its gut, protecting them from disinfection (Caldwell et al., 2003) and even transmitting them vertically to its progeny (Kenney et al., 2005), thus acting as an important potential vector for dispersion (Kenney et al., 2006). Protozoa like *Tetrahymena sp.* (Gourabathini et al., 2008; Rehfuss et al., 2011) but also fungi (Brandl, 2006) have also been shown to interact with *E. coli* and *Salmonella*, and could contribute to their spread in agricultural fields. Finally, important vectors for contamination by *E. coli* may be the plants themselves. Seeds were observed to be colonised during *in vitro* contamination of roots and shoots by *E. coli* and *Salmonella* (Cooley et al., 2003). EHEC could attach and persist on seed surfaces and germinating seedlings (Jeter and Matthysse, 2005) and much effort has been concentrated to effectively decontaminate at this stage of food processing (Taormina and Beuchat, 1999; Beuchat et al., 2001; Scouten and Beuchat, 2002). Studies focusing on seed

colonisation appear to be of prime importance, as the consumption of contaminated sprouted seeds is often linked to food-borne outbreaks of pathogenic *E. coli*.

1.3.1.4. Public health impact of E. coli contamination of plants

Worldwide public demand for "ready-to-eat" produce has increased in the last two decades. Better awareness of food contents and healthy diet requirements in conjunction with public health programs, such as "five-a-day" in the United Kingdom, are believed to contribute to this increase and promote a higher consumption of fruits and vegetables among the population. In the meantime, food-borne infections associated with fresh vegetables are regularly reported (Brandl, 2006; Heaton and Jones, 2008; Mandrell, 2009). The two largest food-borne outbreaks ever observed to be caused by pathogenic *E. coli* in humans occurred in Sakai (Japan) in 1996 (Itoh et al., 1998; Fukushima et al., 1999) and in Germany in 2011 (Rohde et al., 2011), and both involved the association of *E. coli* with vegetable sprouts. Most of the time, EHEC strains of serovar O157:H7 (and rarely *Shigella*) are involved with food-borne diseases linked to *E. coli* (Brandl, 2006; Heaton and Jones, 2008; Mandrell, 2009) (**Figure 1.6**).

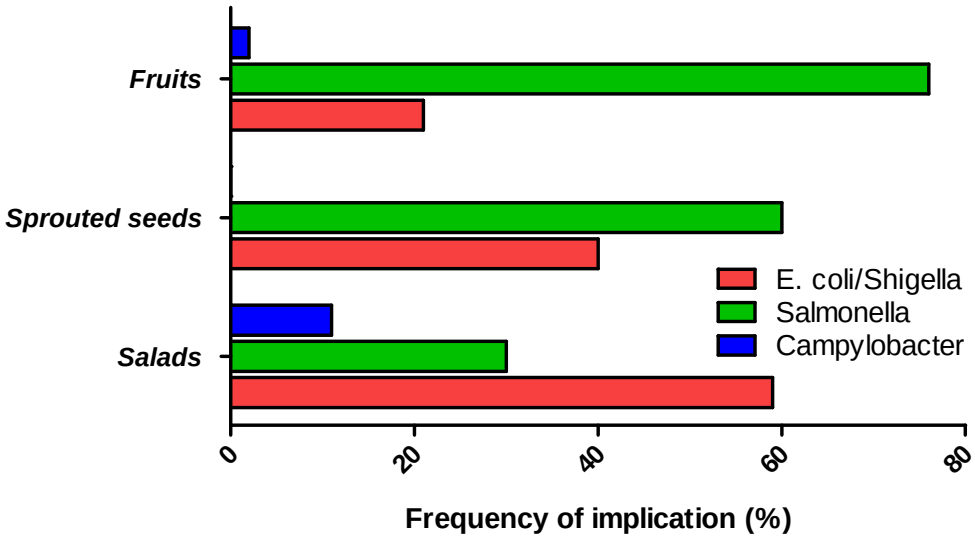


Figure 1.6. Frequency of outbreaks linked to fresh produce in USA between 1990-2004 (Brandl, 2006)**.** The graph shows the different frequencies of implication of a given pathogen in outbreaks linked to the consumption of a given product.

Between 1990 and 2004, about 90 outbreaks of enteric bacterial infections linked to consumption of fruits, salads or sprouted seeds (**Figure 1.6**) have been reported in the United States, most being *Salmonella enterica* and EHEC infections (Brandl, 2006). Additionally, multiple outbreaks of *Salmonella* in UK and USA were linked to consumption of raw tomatoes, cantaloupe melons, strawberries and sprouts (Hedberg et al., 1999; Beuchat, 2002). EHEC strains of serovar O157:H7 were shown to be transmitted by fresh produce in 21% of the food-borne outbreaks from 1982 to 2002 in USA (Rangel et al., 2005). *Salmonella* is more often involved in outbreaks implicating the consumption of fruits and sprouts, whereas *E. coli* is the leading cause of salad-associated outbreaks (**Figure 1.6**). *Campylobacter* has also been sporadically linked to vegetable-related outbreaks (Brandl, 2006).

In 1996, an EHEC strain associated with radish sprouts contaminated the meals of schoolchildren in Sakai, Japan resulting in the death of 3 persons and thousands of illnesses (Fukushima et al., 1999). The strain, named “Sakai” and used for microarray work in this thesis, was subsequently shown to be able to internalise radish tissues (Itoh et al., 1998). Another strain of EHEC O157:H7 caused the 2006 spinach outbreak in USA, during which 3 persons died and 203 became ill (Calvin et al., 2009). The economic impact of the 2006 outbreak was important, as it was the first multistate outbreak causing death in the US linked to food products otherwise perceived by consumers as “healthy” and it was largely covered by national media. More recently, a German outbreak in 2011 added a new, rarer O104:H4 variant to the list of vegetable-related outbreak strains. The O104:H4 strain genome was sequenced within days of the outbreak (Rohde et al., 2011) and revealed a pathogenic mosaicism of both EHEC and EAEC traits, possibly explaining the surprisingly high virulence of the strain (the *E. coli* O104:H4 outbreak caused 50 deaths, 908 HUS cases and 3,167 non-HUS cases and is the deadliest outbreak of *E. coli* ever documented). In Chapter 4 of this thesis, we present results of metabolic profiling using this O104:H4 strain.

Since the 1996 outbreak and exponentially since the 2006 outbreak, investigations on how EHEC could persist, attach, colonise, invade plants and interact with their resident microflora were carried out. In the next section, we present a brief overview of some molecular mechanisms that have been identified to be important for the ecology of pathogenic and non-pathogenic *E. coli* on plants.

1.3.2. Specific association of *E. coli* with leaf surfaces

At the molecular level, much has been done on the investigation of the required functions for attachment and early colonisation of plant surfaces. In this part, we will briefly review these studies with an emphasis on the role of the extracellular matrix and biofilm formation, the role of flagellar motility, internalisation within plant tissues and metabolism.

1.3.2.1. Role of the extracellular matrix in plant attachment and persistence

As mentioned earlier, White et al. (2008) have shown that multicellular behaviour via the production of an extracellular matrix does not promote virulence or intestinal fitness in enteric bacteria. When mice are infected with *Salmonella*, wild-type strains are out-competed by isogenic mutants unable to produce an extracellular matrix, suggesting that aggregation via the "red, dry and rough" (rdar) morphotype is not essential during host colonisation (White et al., 2008). The authors suggest that the primary role of the extracellular matrix is to enhance *Salmonella* survival outside the host, thereby aiding in bacterial dissemination (White et al., 2008). This hypothesis is strongly supported by recent studies showing that curli and cellulose were among other factors that were required for long-term desiccation survival *in vitro* (Gibson et al., 2006; White et al., 2006). A complex regulatory interplay occurs at the *csg* locus encoding curli fimbriae, involving different regulatory networks triggered by various environmental stresses, thereby adding even more complexity to the understanding of multicellular behaviour regulation in enteric bacteria. The optimal conditions for curli

and cellulose expression were established by extensive studies on *E. coli* and *Salmonella* (Vidal et al., 1998; Landini et al., 2006). The general idea that conditions suitable for matrix production seem to be more often present outside the digestive tract has led to the hypothesis that it could be one of the most important factor of extra-intestinal adhesion and survival in natural environments (Olsen et al., 1993), and this role was examined for persistence on plants in *E. coli* and *Salmonella*. In the next two paragraphs, the specific roles of curli and cellulose, the two major components of the extracellular matrix, will be detailed.

In an elegant experiment using attachment-specific function disruption by random transposon insertions, it was shown that production of curli fibres was an important factor for *Salmonella* attachment. Mutants having a transposon insertion in the regulatory region of the *csg* operon did not produce curli or cellulose and showed a 10-fold reduction of attached cell numbers on alfalfa sprouts 4 hours post-infection (hpi) (Barak et al., 2005). In a previous study, the same authors had shown different abilities of enterohemorrhagic strains of *E. coli* (EHEC) and *Salmonella* to attach to alfalfa sprouts (Barak et al., 2002). In this study, EHEC strains could be easily washed away whereas *Salmonella* remained strongly attached (Barak et al., 2002). Most EHEC strains have a point mutation in the *csg* regulatory sequence leading to the absence of curli and cellulose synthesis (Uhlich et al., 2001). When attachment to alfalfa sprouts was tested, curli-producing *E. coli* strains attached as well as *Salmonella*, suggesting bacteria that can produce curli are more likely to attach at high populations to plant surfaces (Barak et al., 2005). The laboratory-adapted strain *E. coli* K-12 has not been observed to attach to tissues (Dong et al., 2003; Matthysse et al., 2005; Torres et al., 2005). However, two studies showed that the addition of

plasmids expressing *csg* genes restored the ability of K-12 to bind alfalfa sprouts, suggesting once again the important role of curli fimbriae in the early attachment to plant tissues (Dong et al., 2003; Matthysse et al., 2005; Torres et al., 2005).

Cellulose biosynthesis (*bcs* genes, regulated by AdrA) was shown to contribute to bacterial attachment to and colonisation of plants (Barak et al., 2007). *bcs* mutants of *Salmonella enterica* were unable to form biofilms at the air/liquid interface, and their ability to attach (4 hours post infection, or "hpi") and colonise (24 and 48 hpi) alfalfa sprouts was drastically diminished (Barak et al., 2007). Using transcriptional GFP fusions and quantitative RT-PCR, authors showed that cellulose biosynthesis pathway genes *adrA* and *bcsA*, and the cellulose and matrix regulator gene *csgD* were upregulated during colonisation of alfalfa sprouts at 48 hpi (Barak et al., 2007), indicating that these functions are important for colonisation of plant surfaces. The same results were obtained for *E. coli* O157:H7 (Matthysse et al., 2008). A deletion mutant analysis indicated that cellulose and poly-β-1,6-N-acetyl-D-glucosamine (PGA) were required for binding to alfalfa sprouts (Matthysse et al., 2008). Moreover, when expressed in *E. coli* K-12 (which normally does not attach to plant tissues), a cellulose synthase from *Agrobacterium tumefaciens* caused a 100-fold increase in the ability to attach to sprouts (Matthysse et al., 2008). However, induction of the synthesis of an exogenous PGA-like polymer originating from *Bordetella bronchiseptica* did not produced clear improvements for attachment of *E. coli* K-12 to alfalfa sprouts (Matthysse et al., 2008). This interesting result suggests that polysaccharides involved in *E. coli* interaction with sprouts may play a very precise and species-specific role rather than being redundant or "accidentally" useful.

Numerous studies have observed that leaf age is an important factor for bacterial colonisation (Brandl and Amundson, 2008). It has been shown that nitrogen availability varied at the surface of lettuce leaves according to leaf age and that it could be a limiting factor for *E. coli* O157:H7 growth. Young-aged lettuce leaves do no leach as much nitrogen as middle-aged leaves (Brandl and Amundson, 2008). Interestingly, it was shown that nitrogen starvation, along with other nutrients was a signal switching on the *csgD* promoter, and thus curli and extracellular matrix expression (Gerstel and Romling, 2001). Conceivably, CsgD-activated structures, among them curli and the extracellular matrix could be, in some cases, useful during persistence in the phyllosphere until the level of nutrients required for growth increased with leaf senescence-dependent leachates.

1.3.2.2. Differences in plant attachment mechanisms between pathogens and commensals

Differences in attachment to plants have been found between pathogenic and non-pathogenic *E. coli* strains, possibly suggesting different strategies of colonisation (Jeter and Matthysse, 2005; Matthysse et al., 2005; Torres et al., 2005). Attachment to alfalfa by curli-deficient Δ*csg* mutants in *E. coli* O157:H7 was not reduced compared to wild-type, whereas attachment of *E. coli* K-12 strains was not possible. However, when *csg* genes were overexpressed in *E. coli* K-12, attachment to alfalfa was observed at high levels (Torres et al., 2005) suggesting that curli could be sufficient to promote binding to plants but that redundant systems may exist in *E. coli* O157:H7 allowing bacteria to compensate for the loss of curli and to retain the ability to bind to plant. Nevertheless, despite the wide distribution of *csg* genes in *E. coli*, curli are not

expressed in many EHEC because of point mutations in the *csgD* promoter (Uhlich et al., 2001; Uhlich et al., 2008). In some *E. coli* K-12 strains, curli are also not expressed because of an amber mutation in the *rpoS* gene (Landini et al., 2006), but expression of *csgD* from a plasmid (Prigent-Combaret et al., 2000) or in an *ompR234* curli-overexpressing background (Vidal et al., 1998) can induce their synthesis. These observations suggest that the use of *E. coli* K-12 and other non-pathogenic strains as surrogates for the study of EHEC attachment on plants should be considered very carefully, as biological mechanisms involved in both cases are very likely to be different (Barak et al., 2002). Additionally, despite interesting results, the use of the laboratory-adapted *E. coli* K-12 strain itself is a major criticism of most of the studies presented here. This unrealistic choice may lead to unconvincing ecological conclusions, as do in general all single-strain mechanistic studies.

Interestingly, features that are only associated with EHEC and O26 pathogens, such as the LEE-encoded EspA filaments synthesised by EHEC serotypes, were required for attachment to salad leaves. Deletion mutants in *espA* were unable to bind rocket salad, lettuce and spinach leaves whereas trans-complementation fully restored the colonisation ability (Shaw et al., 2008). No adherent Δ*escN* mutants were observed. EspB was shown to be important for tropism toward stomata by EHEC, possibly by actively recognising a specific stomatal receptor. Tropism, whose mechanisms remain to be elucidated, was not observed with *E. coli* O26 cells. Effector translocation was not observed with both serotypes, which is coherent with the physical properties of the plant cell wall, which is unlikely to allow protein translocation via the EHEC T3SS (Shaw et al., 2008). This assumption was contrasted by the observation that a EHEC mutant defective in the T3SS ATPase EspN, and thus unable of protein

translocation, was strongly impaired in its spinach and lettuce leaves colonisation abilities (Xicohtencatl-Cortes et al., 2009).

Different pathogen-specific attachment mechanisms have been investigated in various *E. coli* pathotypes for their role in plant attachment. In EHEC, a *fliC* mutant unable to produce its flagellum showed no attachment to spinach and lettuce compared to a motile wild-type strain (Xicohtencatl-Cortes et al., 2009). Similar observations were made in ETEC, in which a *fliC* mutant, but not mutations in *etpA* or *cfa*, which are additional ETEC-associated attachment functions normally involved in the pathogenesis process in animal hosts, showed a significant reduction in plant attachment (Shaw et al., 2011). Conversely, no effect of the flagella was observed in EAEC plant attachment, but a reduction in tropism to stomata was observed (Berger et al., 2009). Additionally, EAEC-specific AAF pili, normally involved in the adhesion of EAEC to mucosa *ex vivo* (Czeczulin et al., 1997) were involved in plant attachment (Berger et al., 2009). Similarly to the variation observed in EAEC and EHEC in *Salmonella*, the role of the flagella in leaf attachment was observed to be serovar-dependent (Berger et al., 2009). A *fliC* mutant of *S. enterica* serovar Senftenberg, involved in a basil-associated outbreak, was impaired in its attachment to basil leaves whereas serovar Typhimurium was not (Berger et al., 2009). It has been shown otherwise that Typhimurium flagella have a role in the active chemotaxis-dependent internalisation of *Salmonella* into lettuce leaves (Kroupitski et al., 2009), suggesting that there are multiple layers of flagellar interaction with the surface of plants.

It is somehow puzzling that a system as specialised as the T3SS in EHEC has just an ectopic attachment role, seemingly without any protein translocation involved. The fact that a mutant in *espN* is impaired in plant attachment would tend to confirm this but on the other hand, contaminating EHEC have not been observed to cause any symptom associated with plant disease or PAMP-mediated bacterial recognition by plants (Zipfel, 2008), suggesting that EHEC and *E. coli* in general are only opportunistic epiphytes. On this latter point, it would seem that *E. coli* is different from *Salmonella*, which has been observed to cause chlorosis and disease symptoms when infiltrated into *A. thaliana* (Schikora et al., 2008; Berger et al., 2011).

These remarkable recent results demonstrate that there are mechanisms involved in EHEC attachment to plants that cannot be studied using non-pathogenic strains lacking a functional T3SS and the LEE region. This section constitutes an interesting indication that mechanisms of early attachment to plants within the *E. coli* species are very diverse.

1.3.2.3. Endophytic lifestyle of E. coli

Since the observation that enteric bacteria could still be isolated on agar medium after decontamination of the leaf surface, many studies have reported the internalisation of enteric bacteria *Salmonella* or *E. coli* in the inner parts of plant leaves. *Salmonella* was found to internalise in tomatoes by stem inoculation (Guo et al., 2001), in sprouting mung beans (Warriner et al., 2003), parsley leaves (Lapidot et al., 2006), barley (Kutter et al., 2006) and *Arabidopsis thaliana* (Cooley et al., 2003; Schikora et al., 2008). *E. coli* (not only pathogenic strains) was also supposedly able to actively

reach the apoplast in lettuce and *A. thaliana* after soil contamination (Wachtel et al., 2002; Cooley et al., 2003), spinach (Warriner et al., 2003) and mung bean sprouts (Warriner et al., 2003). Root inoculation with both *E. coli* and *Salmonella* resulted in shoot contamination and this was found to be dependent on flagellar motility (Cooley et al., 2003). Also, *Salmonella* mutants in chemotaxis systems lost the ability to internalise in lettuce, interestingly in a light-dependent process (Kroupitski et al., 2009). Many plant sites were observed to be internalised by *E. coli* and *Salmonella* (**Figure 1.7**), with a preference in leaves and vascular tissues, presumably richer in nutrients.

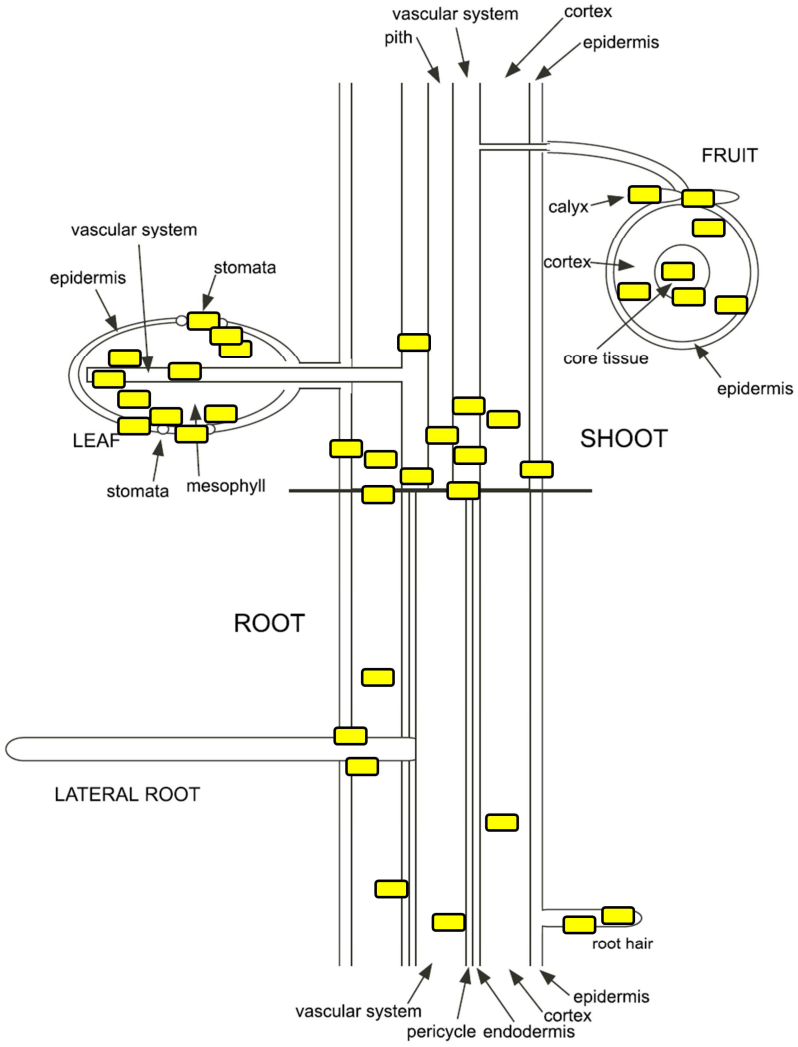


Figure 1.7. Schematic representation of the different plant sites in which internalisation of *E. coli* or *Salmonella* has been observed (Deering et al., 2011). The yellow shapes represent bacteria, not to scale. This figure was adapted from a meta-analysis study and is copyrighted by Elsevier Ltd.

The experimental proof of internalisation is not easy to achieve and has produced conflicting results in many studies as many factors can presumably affect the outcome and reproducibility of internalisation experiments, such as the age and species of plants or the strains used (Warriner and Namvar, 2010). Protocols of infection also seem to play a role, as the internalisation of the same strain of *E. coli* in spinach was only observed when water was used to inoculate leaves and not via soil and roots (Mitra et al., 2009). A recently published meta-analysis thoroughly listed all studies

examining the internalisation of *E. coli* or *Salmonella* into plants, and estimated that internalisation was observed with a 70% success rate, depending on protocol (Deering et al., 2011). Nevertheless, the use of fluorescent reporters and microscopy has produced convincing evidence (Schikora et al., 2008; Kroupitski et al., 2009) that under certain strain-specific and environmental circumstances, *E. coli* and *Salmonella* can internalise into plants and thus potentially evade surface decontamination techniques.

1.3.2.4. Metabolic opportunities for E. coli *growth on plants*

Plant leaves carry a wide variety of inorganic and organic compounds that are leached or exuded from the internal tissue (Morgan and Tukey, 1964) and that can act as nutrient sources to bacteria on leaves. Consumption of nutrients from a leaf surface by bacteria has been demonstrated using radio-labelled carbonated sources (Rodger and Blakeman, 1984). It has also been observed that the structure of bacterial populations on leaves could be manipulated by changing nutrient availability on the plant surface (Wilson and Lindow, 1994; Wilson et al., 1995), indicating that microbes growing on plant surfaces could be competing for a very limited amount of nutrients, which in turn, would determine the structure and size of microbial populations. Depending on leaf age, growing conditions and the presence of wounds, various amounts of carbohydrates, organic acids, amino acids, methanol and various salts are available for bacteria on or within leaves (Mercier and Lindow, 2000). Several studies to assess nutrient, water or salts availability on leaves have been conducted using reporter biosensor strains. Sugar availability and localisation on leaves has been investigated using an engineered strain of *Erwinia herbicola*

expressing a green fluorescent protein (GFP) under the control of the fructose- and sucrose-responsive promoters P_{fruB} and P_{scrY} (Leveau and Lindow, 2001). Microscopic observations revealed that bacterial consumption of sugars was generally in localised sites on the plant rather than randomly dispersed across the leaf, which suggests that most areas of a leaf harbour only small amounts of nutrients, and that nutrients may be abundant in only a few locations (Leveau and Lindow, 2001). In the same way, water availability on a leaf has been assessed using strains of *Pantoea agglomerans* expressing a green fluorescent protein (GFP) reporter gene linked to a promoter (P_{proU}) responsive to water availability (Lindow and Brandl, 2003). Most cells do not experience the expected water stress on dry leaves, suggesting that bacteria can survive using the desiccation-resistant aqueous thin layer surrounding leaves whose function is to retain moisture emitted through stomata (Lindow and Brandl, 2003). Ferric iron availability has also been assessed using biosensor strains. Using the Fe^{3+}-sensitive regulation of the promoter P_{pvd} controlling the expression of a siderophore membrane receptor in *Pseudomonas syringae*, a spatial heterogeneity of available iron on plant surfaces was observed, even if iron is apparently not limiting bacterial growth of bacteria on leaves (Joyner and Lindow, 2000).

Nitrogen availability, however, was variable at the surface of young, medium-aged and old lettuce leaves and was found to be a limiting factor for growth of *E. coli* O157:H7 on leaves (Brandl and Amundson, 2008). On lettuce leaves, nutrients may leach by guttation (i.e., the exudation of water from leaves by hydathodes as a result of root pressure) which is considered to be a common phenomenon in this plant. Such fluids contain diverse substances, including amino acids, carbohydrates and inorganic substances like NO_3^+ and NH_4^+ ions, which may contribute to nitrogen availability on

the leaf surface. It was suggested that guttation fluids were contributing to high levels of nitrogen availability on young lettuce leaves, providing good conditions for bacterial immigration and persistence when the crop is young (Brandl and Amundson, 2008). The role of trichomes as nutrient sources has been hypothesised, as these structures are believed to leach nutrients to the plant microflora (Morris and Monier, 2003; Monier and Lindow, 2004). Indeed, microscopic studies have shown that epiphytic bacteria and *E. coli* O157:H7 were more likely to form aggregates near the base of glandular trichomes (Morris and Monier, 2003; Monier and Lindow, 2004; Brandl and Amundson, 2008). However, the presence and amount of hydathodes and trichomes on the leaf surface differs among species, suggesting that different bacterial colonisation strategies could exist for different types of plants.

Research from naturally plant-associated bacteria or phytopathogens can provide clues on what the conditions are at the bacterial scale of *E. coli* when immigrating onto a leaf. Numerous studies on plant roots have highlighted what could be expected in the phyllosphere regarding the role of exudates, as root and leaf exudates are both composed of a complex mixture of sugars, amino acids, organic acids and other compounds (Mercier and Lindow, 2000; Lugtenberg et al., 2001). Metabolic versatility seems to be an advantage in both rhizosphere and phyllosphere colonisation (Lugtenberg et al., 2001). Mutations affecting the metabolism of sugars or organic acids of bacteria resulted in a decreased fitness in the rhizosphere (Lugtenberg et al., 2001). Synthesis of sugar transport and catabolic enzymes are also enhanced in response to root exudates in the rhizosphere (Espinosa-Urgel, 2004), thus mutants in these proteins are believed to be poorer colonisers (Lugtenberg et al., 2001; Espinosa-Urgel, 2004). Metabolic activity of *Pseudomonas sp.* seems to vary

according to the position on roots, presumably because of differences in O_2 solubility and exudate composition in the different parts of the root (Kragelund et al., 1997).

As presented in this non-exhaustive review of the literature, the mechanisms of short-term colonisation of plants by *E. coli* involving attachment and early persistence are starting to be very well understood. It is equally interesting to understand how *E. coli* persists on plants and generally in secondary environments in the longer term. The identification of plant- or nonhost-specific metabolic abilities or associated behaviour, if any, is part of this effort, and may give important knowledge in defining the roles of secondary environments in the ecology of *E. coli*, and the possibility to control problematic pathogens more efficiently.

1.4. Context of this work

As presented in this introduction, the general evolution and ecology of *E. coli* as well as its mechanisms of plant association are extensively studied. Regarding this last topic, because of the economic pressure on finding new ways to control human pathogens on plants and more generally on food, regular updates on the advancement are published and the need for more research highlighted (Niemira et al., 2009; Teplitski et al., 2009). In this work, we chose to address topics that are relevant academically (the ecology of *E. coli*) and industrially (the improvement of food safety and monitoring):

- **Where does *E. coli* contaminating agricultural fields generally come from?**

Microbial source tracking has not proven very successful for *E. coli*. Given the mosaic nature of *E. coli* genomes, it is very hard to identify genotypes of markers that are specific to certain hosts with certainty. A new approach can be considered by trying to find population-associated traits in various environments. In the Introduction, we showed that different environments shape different population structures in *E. coli*. In that context, it would be interesting to identify the population structure in a sampled collection of *E. coli* isolates from plants, as no other study to our knowledge has sought to characterise such a collection. This topic is addressed in Chapter 3.

- **Are there specific functions or traits in plant-associated *E. coli*?**

Many molecular mechanisms involved in attachment, persistence and colonisation of plants and soil have been demonstrated using single-strain studies, as introduced

above. In this work, we adopt a population-wide approach by characterising a large collection of *E. coli* from plants. Conceivably, some of the traits that were found to be involved in the interaction could be over-represented in plant-associated isolates, confirming their true ecological importance. Additionally, if such traits are identified, it would be plausible to develop a way to quantify this "adaptation" to plants, or more generally nonhost environments, for single *E. coli* strains, in order to simply predict their likelihood of plant contamination. This part was developed in Chapter 4.

- **Does *E. coli* interact with the plant resident microflora?**

The primary environment of *E. coli* is densely populated with the gut microflora, and resident or transient *E. coli* strains are in constant direct or indirect interaction with some of its members. Similarly, many diverse microbes colonise the plant environments, from soils and roots to shoots. There are mechanisms of colonisation resistance in the gut, for which the resident microflora can affect the colonisation outcome of foreign and potentially pathogenic microbes, and it is interesting to examine how the plant resident microflora reacts when exogenous *E. coli* are colonising. This subject is addressed in a preliminary way in Chapter 5.

2. Experimental procedures

2.1. Bacterial strains and isolation of environmental *E. coli*

2.1.1. Phyllosphere isolates

2.1.1.1. Microbiological methods of isolation

In this work, we gathered 106 *E. coli* strains isolated from the aerial parts of field-grown plants into a collection called "GMB". These strains were isolated from various salad crops, mostly growing in agricultural fields in England, between summer 2008 and autumn 2009. Most isolates were retrieved from agricultural fields where plants were grown to be commercialised as food. Strains GMB01 to GMB58, and GMB87 to GMB110 were isolated from post-harvest material in a partner food safety inspection laboratory. All strains were sent to us on agar stabs. Selective isolation of *E. coli* isolates from plants was performed as recommended by the BS ISO 16649:2001 standard procedure, which involves stomaching (i.e. crushing and stirring leaf material) in a recovery diluent, which is typically a derivative of peptone buffer saline (PBS) diluent. Bacterial suspensions in PBS are then filtered to remove plant debris, diluted, plated on tryptone bile X-glucuronide (TBX) agar plates and incubated at 44°C for 18-24 hours. Growth on TBX agar is selective for bacteria able to survive the presence of bile salts, which are typically inhibiting Gram-positive bacteria. The addition of 5-bromo-4-chloro-3-indolyl-beta-D-glucuronide (or X-glucuronide) in the medium provides a chromogenic identification of bacteria possessing the X-glucuronidase enzyme encoded by the *uidA* gene. *E. coli* strains

generally possess *uidA*, and are able to cleave X-glucuronide, resulting in blue colonies on TBX agar. There can be false negatives on TBX as a small fraction of *E. coli*, notably *E. coli* O157:H7 and some *Shigella* do not possess *uidA* and appear as white colonies on TBX. Also, it has been reported that the *uidA* gene could be present but the enzyme not expressed (Martins et al., 1993) possibly because of mutations in the promoter region.

Strains GMB59 to GMB86 were isolated by Jeanette Newman (IFR) and Lucile Mayeux (IFR) in September 2008 from a field near Martham, Norfolk (UK) according to the same standard procedures. Upon retrieval of organisms, turbidity and gas production on brilliant green bile broth containing 2% of lactose (BGLBB; cat: CM0031; Oxoid Ltd.) was tested and all isolates were positive for both, which when coupled with the growth phenotype on TBX, confirmed that the biochemically tested isolates were *E. coli*.

2.1.1.2. Geographical, temporal and plant origin of isolation

GMB strains were isolated from plants grown in various agricultural fields located mainly in 5 different locations in England and in Italy, from the aerial parts of various plants (**Figure 2.1**).

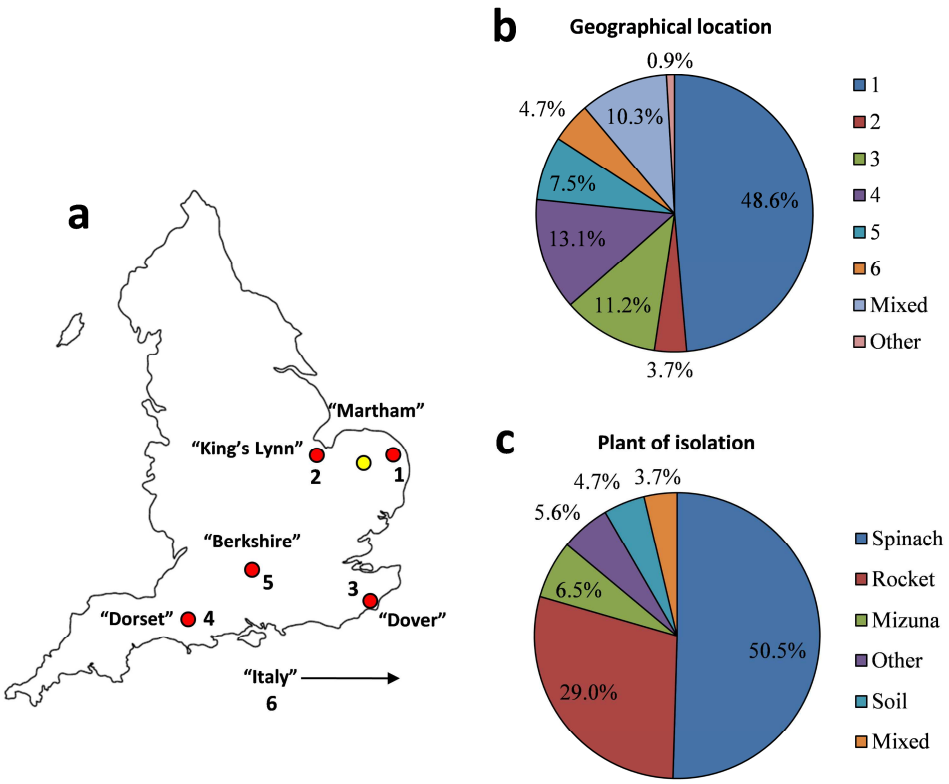


Figure 2.1. Information on isolation of GMB strains. (a) Geographical location of isolation, (b) proportion of isolates from these geographical locations and (c) plant of isolation. Red dots on the map indicate sampling sites, the yellow dot indicates the location of our laboratory.

GMB strains were mostly associated with spinach (*Spinacia oleracea*) and rocket (*Eruca sativa*), and to a lesser extent with other types of salad that included mizuna (common name for *Brassica rapa nipposinica* and *B. juncea var. japonica*), tatsoi (*Brassica narinosa*), amaranth leaves (*Amaranthus sp.*), red chard (*Beta vulgaris*) and watercress (*Nasturtium officinale*). GMB82 to GMB86 were isolated from the soil of a rocket field located around Martham, Norfolk, UK in September 2008. GMB103 has a different isolation history from the rest, as it was isolated from corn (*Zea mays*) grown in south-eastern Asia. GMB strains were mostly isolated from plants grown during summer and autumn 2008 (n=84). Twenty-two strains were isolated in 2009. It

is not known when the asian corn plant colonised by GMB103 was growing. A full description with strain names can be found in **Table 2.1**.

Table 2.1. Individual GMB strain information.

Strain name	Plant of isolation	Location of isolation	Date of isolation
GMB01	Rocket	Italy	June 2008
GMB02	Rocket	Italy	June 2008
GMB03	Rocket	King's Lynn	July 2008
GMB04	Rocket	King's Lynn	July 2008
GMB05	Mizuna	King's Lynn	July 2008
GMB06	Spinach	Berkshire	July 2008
GMB07	Spinach	Dover	July 2008
GMB08	Spinach	Martham	July 2008
GMB09	Spinach	Martham	July 2008
GMB10	Spinach	Martham	July 2008
GMB13	Spinach	Berkshire	July 2008
GMB14	Mizuna	King's Lynn	July 2008
GMB15	Spinach	Martham	July 2008
GMB16	Mizuna	Martham	July 2008
GMB17	Mizuna	Martham	July 2008
GMB18	Spinach	Dorset	July 2008
GMB19	Other	Dorset	July 2008
GMB20	Rocket	Martham	July 2008
GMB21	Spinach	Martham	July 2008
GMB22	Spinach	Dorset	July 2008
GMB23	Spinach	Dover	July 2008
GMB24	Spinach	Martham	July 2008
GMB25	Spinach	Martham	July 2008
GMB26	Other	Dorset	July 2008
GMB27	Mizuna	Martham	July 2008
GMB28	Spinach	Berkshire	July 2008
GMB29	Mixed	Martham	August 2008
GMB30	Spinach	Dover	August 2008
GMB31	Spinach	Dover	August 2008
GMB32	Spinach	Dorset	August 2008
GMB33	Spinach	Dorset	August 2008
GMB34	Spinach	Dorset	August 2008

GMB35	Other	Martham	August 2008
GMB36	Mizuna	Dorset	August 2008
GMB37	Spinach	Dover	August 2008
GMB38	Spinach	Berkshire	August 2008
GMB39	Spinach	Dorset	August 2008
GMB40	Spinach	Dover	August 2008
GMB41	Other	Dorset	August 2008
GMB42	Spinach	Berkshire	August 2008
GMB43	Spinach	Dorset	August 2008
GMB44	Mixed	Mixed	August 2008
GMB45	Spinach	Mixed	August 2008
GMB46	Spinach	Mixed	August 2008
GMB47	Spinach	Dover	August 2008
GMB48	Spinach	Martham	August 2008
GMB49	Spinach	Martham	August 2008
GMB50	Spinach	Berkshire	August 2008
GMB51	Spinach	Berkshire	August 2008
GMB52	Spinach	Dover	August 2008
GMB53	Spinach	Martham	August 2008
GMB54	Spinach	Martham	August 2008
GMB56	Spinach	Mixed	August 2008
GMB57	Spinach	Mixed	August 2008
GMB58	Other	Mixed	August 2008
GMB59	Rocket	Martham	September 2008
GMB60	Rocket	Martham	September 2008
GMB61	Rocket	Martham	September 2008
GMB62	Rocket	Martham	September 2008
GMB63	Rocket	Martham	September 2008
GMB64	Rocket	Martham	September 2008
GMB65	Rocket	Martham	September 2008
GMB66	Rocket	Martham	September 2008
GMB67	Rocket	Martham	September 2008
GMB68	Rocket	Martham	September 2008
GMB69	Rocket	Martham	September 2008
GMB70	Rocket	Martham	September 2008
GMB71	Rocket	Martham	September 2008
GMB72	Rocket	Martham	September 2008
GMB73	Rocket	Martham	September 2008
GMB74	Rocket	Martham	September 2008
GMB75	Rocket	Martham	September 2008

GMB76	Rocket	Martham	September 2008
GMB77	Rocket	Martham	September 2008
GMB78	Rocket	Martham	September 2008
GMB79	Rocket	Martham	September 2008
GMB80	Rocket	Martham	September 2008
GMB81	Rocket	Martham	September 2008
GMB82	Soil	Martham	September 2008
GMB83	Soil	Martham	September 2008
GMB84	Soil	Martham	September 2008
GMB85	Soil	Martham	September 2008
GMB86	Soil	Martham	September 2008
GMB87	Spinach	Dover	August 2009
GMB88	Spinach	Martham	September 2009
GMB89	Spinach	Martham	August 2009
GMB90	Spinach	Martham	August 2009
GMB91	Spinach	Mixed	September 2009
GMB92	Rocket	Dorset	August 2009
GMB93	Spinach	Martham	October 2009
GMB94	Spinach	Dover	September 2009
GMB95	Spinach	Dover	August 2009
GMB96	Spinach	Italy	October 2009
GMB97	Spinach	Dover	August 2009
GMB98	Spinach	Dorset	September 2009
GMB99	Rocket	Italy	August 2009
GMB100	Spinach	Mixed	August 2009
GMB101	Mixed	Mixed	August 2009
GMB102	Rocket	Dorset	August 2009
GMB103	Other	Other	Unknown
GMB104	Spinach	Martham	September 2009
GMB105	Spinach	Italy	October 2009
GMB106	Spinach	Martham	August 2009
GMB107	Mixed	Mixed	September 2009
GMB108	Spinach	Martham	August 2009
GMB110	Spinach	Mixed	August 2009

2.1.2. *E. coli* reference (ECOR) collection

In this work, we compared strains of *E. coli* with various isolation histories to examine the impact of secondary environments on selected phenotypes and genotypes. To represent strains from the primary environment of *E. coli*, I chose the 72 members of the well-described *E. coli* reference (ECOR) collection. The ECOR collection was assembled in 1983 from a larger collection of 2,600 strains for the purpose of gathering strains representing the whole genotypic variability in the *E. coli* species, as based on variability in multilocus enzyme electrophoretic (MLEE) profiles (Ochman and Selander, 1984). ECOR strains were all originally collected by Roger Milkman from faecal samples of healthy zoo animal or humans, with the exception of 10 ECOR strains that were isolated from the urine of women with urinary tract infections (acute pyelonephritis or acute cystitis). One strain was isolated from a woman with asymptomatic bacteriuria (ECOR71).

2.1.3. Other strains used in this work

For metabolic profiling using BIOLOG microplates, we used 2 reference strains: *E. coli* K-12 strain MG1655 and a Shiga toxin-deficient deletion mutant of *E. coli* O157:H7 strain Sakai. Both strains were kindly provided by Martin Goldberg (University of Birmingham, UK). We characterised siderophore production of various pathogenic strains, including 3 other Shiga-toxin deletion mutants of O157:H7 isolates from food which were kindly given by Kirrilly Wilson (IFR). In the same part, we also used 10 APEC strains, kindly sent by Timothy Johnson (University of Minnesota, USA) and the probiotic *E. coli* strain Nissle 1917, kindly given by Ulrich

Sonnenborn (Ardeypharm GmbH, Germany). Finally, we are grateful to Dr Geraldine Smith (Health Protection Agency Colindale, London, UK) for sending our colleague Dr Stephanie Schüller (IFR) the 2011 O104:H4 strain (isolate H1 1218 0280), of which she kindly performed a metabolic profiling experiment in the IFR category-3 laboratory.

2.1.4. Long-term storage of strains

Strains were all stocked using the same procedure, to ensure a very high concentration of bacteria in the stock collection tube. A pure culture was plated on agar medium (LB or TBX) and incubated for 24 h at 37°C. After incubation, 3 ml of LB were poured over the plate and bacteria were scrapped using a sterile spreader. One millilitre of the resulting suspension was transferred in 1 ml of 40% glycerol in a cryovial. Frozen stocks were kept at −80°C. For each culture set up from this frozen collection, tubes were preliminarily transferred to dry ice and minimum manipulation was performed.

2.2. BOX-PCR

2.2.1. Principle of the method

It has been found that bacterial genomes contain repetitive DNA sequences that can be located in inter- or intragenic regions (Tobes and Ramos, 2005) and can represent up to 5% of the whole genome (Ussery et al., 2004). The biological role of these repetitive elements remains unknown, but some have hypothesised that it may be

involved in RNA or DNA metabolism (Tobes and Ramos, 2005; Ishii and Sadowsky, 2009). Repetitive sequences in bacterial genomes are often the cause of genome rearrangements, either inversions of genomic regions or deletions of large fragments (see Introduction section 1.1.2 for more details). By using PCR with primers targeting the repetitive elements, one can generate a mixture of amplicons of different sizes that depend on the distance between each repetitive element. Electrophoretic migration of this mixture of amplicons generates a "profile" for any individual strain. The REP-PCR method targets repetitive extragenic palindromic DNA sequences disseminated throughout bacterial genomes (Tobes and Ramos, 2005). Many derivatives of the REP-PCR method were developed (Ishii and Sadowsky, 2009), including BOX-PCR, a single-primer method targeting mosaic repetitive elements called BOX elements (Martin et al., 1992). The principle is summarised below.

Bacterial genomic rearrangements can happen during homologous recombination, when portions of the genome are inserted or deleted after horizontal gene transfer. BOX-PCR is used to examine genomic diversity or clonality between strains, and the relatedness of electrophoretic profiles within a population or a collection of strains (Louws et al., 1994; van Belkum et al., 1996). Two microbiological clones will be very likely to share the same BOX-PCR profile, as no major rearrangements are likely to have occurred. Conversely, one can hypothesize that 2 divergent strains will have different recombination histories. Genomic rearrangements will result in insertions or deletions that influence the BOX amplicons and their sizes, and thus the resulting profiles. It is important to keep in mind that this method is not phylogenetically accurate, as similar profiles from phylogenetically distant strains have already been observed (Lynch, 1988). Nevertheless and despite these approximations, it is

generally safe to use BOX-PCR as a cheap and fast method to distinguish between clones within a known context, for instance during epidemics (Currie et al., 2007), host colonisation studies (Martinez-Medina et al., 2009) or as a general method to observe genetic diversity.

2.2.2. BOX-PCR protocol used in this study

In this study, we used a previously published protocol (Versalovic et al., 1993). ECOR and GMB strains were cultured on LB agar plates overnight at 37°C. BOX-PCR was performed in 75 µl-reactions from fresh colonies using the 2X Go-Taq Green Master Mix (Promega, UK), 5 µl of lysed colony suspension, with addition of a final concentration of 0.8 nM of BOXA1R primer (5'-CTACGGCAAGGCGACGCTGACG-3') (Versalovic et al., 1993) and 0.1 µg/µl of bovine serum albumin (cat: 10711454001; Roche Applied Science). PCR conditions were as follows: 1 cycle at 95°C for 20 min, 30 cycles at 90°C for 30s, 52°C for 1 min and 72°C for 8 min, and 1 cycle at 72°C for 16 min.

Before electrophoresis, an experimental plan was designed to randomly distribute the tested strains in batches of 22 or 23 wells corresponding to samples in a 25-well agarose gel. Two wells at each end and one lane in the middle of each gel were reserved for migration of 2-log DNA Ladder (cat: N3200; New England Biolabs) to facilitate the comparison of profiles across gels. All samples were run on 1.5% TAE-agarose electrophoresis for 1H at 100V on the same day using the same electrophoresis buffer batch to minimize variability when comparing profiles across gels. Pictures of the agarose gels were aligned in a graphical editing program and

analysed using TotalLab Quant/Phoretix software (Nonlinear Dynamics Ltd., http://www.nonlinear.com/). Where isolates (3/181) did not show any bands on an electrophoresis gel, this was attributed to technical problems and the result excluded from the statistical analyses. Background-subtracted pixel intensity data, retention factor (R_f) values and peak height were extracted to a Microsoft Excel spreadsheet, which was processed in MATLAB with help from E. Katherine Kemsley from the Bioinformatics and Statistics department at IFR.

2.2.3. Controls and statistical analysis

Electrophoresis is a technique coupling a timely application of electric current through an agarose gel, thereby potentially causing considerable technical variation between samples. We looked at a potential "gel effect", i.e. an artificially created similarity of samples that are run on the same agarose gel and whose picture is taken at the same time. This phenomenon was observed during the development of this method and its subsequent statistical analysis, which is why we randomised the samples, and made and ran the agarose gels as reproducibly as possible. Defining previously known parameters for our dataset, we performed unsupervised modelling (partial least squares, or PLS modelling) to the 8 different gels, each defined as a distinct parameter for the PLS model in order to look for a potential "gel effect". There were no significant differences between profiles that could be explained by the presence on any particular gel (data not shown).

2.3. *E. coli* phylogroup assignment using a triplex PCR method

2.3.1. Background

Following the observation that almost all *E. coli* strains cluster together in relatively stable phylogenetic clades, major and minor phylogroups were defined. Among all clades, phylogroups A, B1, B2 and D seemed to gather the majority of strains and were considered as major phylogroups. Experimental methods to accurately assign isolates to a phylogroup rely on ribotyping or multilocus techniques such as MLEE or MLST which are costly, complex and/or time-consuming. Clermont et al. (2000) suggested a simple and accurate method to assign *E. coli* strains to major phylogroups A, B1, B2 and D using a multiplex PCR amplification of 3 genetic markers in a way that different combinations of markers are associated with different phylogroups (**Table 2.2**).

Table 2.2. The possible outcomes of a triplex PCR experiment to determine phylogroups. "+" denotes the detection of a band for the corresponding primer pair, and "-" denotes the absence of band.

Phylogroup	ChuA	YjaA	TspE4.C2
A	-	-	-
	-	+	-
B1	-	-	+
B2	+	+	-
	+	+	+
D	+	-	-
	+	-	+

The accuracy of the method has been examined in a few studies. In the original report (Clermont et al., 2000), 230 strains, including 66 ECOR isolates, of which the phylogroup assignment was known by other methods were found to be classified correctly the overwhelming majority. A later report used more strains in a goal to refine accuracy (Gordon et al., 2008). Authors observed that strains assigned to phylogroup D based on the presence of ChuA and TspE4.C2 (or a "+–+" amplification pattern) could be in some cases B2 phylogroup false negatives, and suggested the testing for the presence of the *ibeA* gene to correctly reassign "+–+" isolates to phylogroup B2 (Gordon et al., 2008). A more recent study on the diversity of *E. coli* wild-type isolates from freshwaters in Michigan highlighted the rare possibility of apparent phylogroup transfer based on gain or loss of markers used in the triplex PCR method (Walk et al., 2007).

2.3.2. Protocol used in this study

We used PCR conditions as described previously (Clermont et al., 2000). Primer sequences are detailed in Table 2.3.

Table 2.3. Primer sequences for triplex PCR determination of *E. coli* phylogenetic group.

Name	Sequence (5'>3')	Amplicon length (bp)
ChuA.fw	GACGAACCAACGGTCAGGAT	279
ChuA.rev	TGCCGCCAGTACCAAAGACA	
YjaA.fw	TGAAGTGTCAGGAGACGCTG	211
YjaA.rev	ATGGAGAATGCGTTCCTCAAC	
TspE4.C2.fw	GAGTAATGTCGGGGCATTCA	152
TspE4.C2.rev	CGCGCCAACAAAGTATTACG	

Strains to test were grown on LB agar overnight at 37°C after which a single colony was picked and resuspended in 50 µl of sterile ultrapure water. Two-microliters of each of these suspensions were used as templates for PCR in a 50 µl reaction volume containing 25 µl of the 2X Go-Taq Green Master Mix (cat: M712; Promega) and 20 pmol of each primer diluted in sterile ultrapure water.

Amplification was carried out as follows: 1 cycle at 94°C for 5 min, 30 cycles at 94°C for 30s, 55°C for 30s, 72°C for 30s, and 1 cycle at 72°C for 7 min. Five-microliters of the amplified reaction were used for electrophoresis migration for 45min to 1h on a 1.5% agarose gel (prepared in 1X Tris-acetate-EDTA buffer, or TAE buffer), using 5µl/lane of 2-log DNA ladder (cat: N3200; New England Biolabs) as a size marker.

2.4. Multilocus sequence typing (MLST) and diversity analyses

2.4.1. Principle of MLST

The apparition of a multilocus method like MLEE greatly improved the ability to phylogenetically discriminate between bacteria. Its sequence-based variant MLST added reproducibility and ease of use, and has since been proven to be one of the most phylogenetically discriminating methods so far. It is now widely used in bacterial epidemiology, with public databases covering a few dozen organisms.

The general principle of MLST is to obtain for each tested isolate the sequences of 6 to 8 internal fragments from housekeeping genes, believed not to be under adaptive (or positive) selection. This point is central to MLST, as variations in sequences

evolving in the absence of positive selection are caused either by genetic drift, which likely constitutes a more stable phylogenetic signal, or by purifying selection, which tends to maintain synonymous substitutions. Methods such as the d*N*/d*S* ratio calculation can highlight if specific loci are under positive selection or not by comparing the rates of synonymous and non-synonymous substitutions (Yang and Bielawski, 2000). Each distinct sequence for each housekeeping gene is called an allele type (AT) and the combination of 8 AT constitutes a sequence type (ST). Isolates with identical ST are called "clones" and isolates with similar ST (5, 6 or 7 similar AT) are in the same "clonal complex" (CC). It is worth noting that two MLST clones are not necessarily biological clones, but may be representative of a stable and common sequence type or clonal complex in the tested species. Additionally, this terminology illustrates that MLST is mainly a method to examine bacterial clonality, in order to establish the possibly common evolutionary history of the tested isolates, a feature extremely useful when examining the epidemiology of pathogens.

There are currently 3 available schemes for *E. coli* MLST studies, which are commonly called by their country of origin ("American", "German" and "French"), and use the sequencing of different genes. All schemes generally produce consistent observations on the population structure of *E. coli* (Gordon et al., 2008). An extended scheme suggesting the additional sequencing of up to 22 genes for a single isolate has also been published (Walk et al., 2009).

2.4.2. MLST protocol used in this study

In this work, we used the *E. coli* scheme proposed by Jaureguy et al. (2008) based on the sequencing of internal fragments of 8 housekeeping genes (**Table 2.4**).

Table 2.4. Genes from MLST scheme used in this study and their product.

Name	Gene product
dinB	DNA polymerase
icdA	isocitrate dehydrogenase
pabB	*p*-aminobenzoate synthase
polB	polymerase PolIII
putP	proline permease
trpA	tryptophan synthase unit A
trpB	tryptophan synthase unit B
uidA	β-glucuronidase

The primers detailed in **Table 2.5** were used for amplification (primers oF and oR were used for the sequencing of all fragments).

Table 2.5. Primer sequences used in the MLST experiment.

Gene	Direction	Sequence (5' > 3')
dinB	fw	GTTTTCCCAGTCACGACGTTGTATGAGAGGTGAGCAATGCGTA
	rev	TTGTGAGCGGATAACAATTTCCGTAGCCCCATCGCTTCCAG
icdA	fw	GTTTTCCCAGTCACGACGTTGTAATTCGCTTCCCGGAACATTG
	rev	TTGTGAGCGGATAACAATTTCATGATCGCGTCACCAAAYTC
pabB	fw	GTTTTCCCAGTCACGACGTTGTAAATCCAATATGACCCGCGAG
	rev	TTGTGAGCGGATAACAATTTCGGTTCCAGTTCGTCGATAAT
polB	fw	GTTTTCCCAGTCACGACGTTGTAGGCGGCTATGTGATGGATTC
	rev	TTGTGAGCGGATAACAATTTCGGTTGGCATCAGAAAACGGC
putP	fw	GTTTTCCCAGTCACGACGTTGTACTGTTTAACCCGTGGATTGC
	rev	TTGTGAGCGGATAACAATTTCGCATCGGCCTCGGCAAAGCG
trpA	fw	GTTTTCCCAGTCACGACGTTGTAGCTACGAATCTCTGTTTGCC
	rev	TTGTGAGCGGATAACAATTTCGCTTTCATCGGTTGTACAAA
trpB	fw	GTTTTCCCAGTCACGACGTTGTACACTATATGCTGGGCACCGC
	rev	TTGTGAGCGGATAACAATTTCCCTCGTGCTTTCAAAATATC
uidA	fw	GTTTTCCCAGTCACGACGTTGTACATTACGGCAAAGTGTGGGTCAAT
	rev	TTGTGAGCGGATAACAATTTCCCATCAGCACGTTATCGAATCCTT
oF		GTTTTCCCAGTCACGACGTTGTA
oR		TTGTGAGCGGATAACAATTTC

Reactions were prepared and performed in 96-well PCR plates (Microplate Abgene Thermo-Fast, cat: TUL-962-005Y; Fisher Scientific) as follows: for a 50µl reaction, 25µl of 2X Go*Taq* Colorless Master Mix (Promega), 2µl of 10 µM primer mix (forward and reverse), 18µl of ultrapure water and 5µl of boiled cell lysate. Amplification was performed as follows: 1 cycle at 95°C for 2 min, 30 cycles at 95°C for 30s, 50°C or 55°C for 30s and 72°C for 5 min, and 1 cycle at 72°C for 5 min.

Most of MLST studies also use *Taq* polymerase, because of its very good efficiency for the high-throughput amplification of many templates from boiled lysates (Ibarz Pavon and Maiden, 2009; Walk et al., 2009). However, *Taq* polymerases do not exhibit a 3'-5' exonuclease activity, which allows for the correction of incorrectly incorporated bases during polymerisation, leading to errors at the rate of 1 for each 9000 nucleotides (Tindall and Kunkel, 1988). If an error arises during the first cycles,

it is amplified and can lead to misinterpretation. To circumvent this issue, we halved reaction mixes and performed two independent amplification steps and mixed back the amplicons, in an effort to reduce the frequency of potential mistakes during the first steps of amplification. Using this procedure, we did not subsequently detect any problem potentially linked to the lack of a *Taq* proofreading ability. All amplicons were checked on a 2% agarose gel (prepared in 1X TAE buffer) using 5 µl/lane of 2-log DNA ladder (cat: N3200; New England Biolabs) as a size marker. Samples with no amplification were reamplified with a lower annealing temperature (50°C instead of 55°C). If there was still no amplification, the whole isolate was discarded and its other fragment not sequenced.

Amplicons were purified before sequencing using a high-throughput 96-well plate method. We used the vacuum-based Qiagen MinElute 96 UF PCR purification kit (cat: 28051; Qiagen) according to the manufacturer's recommendations. Briefly, we transferred 50 µl of PCR products (i.e. the mixture of 2 PCR suspensions) in the PCR purification plate, placed the plate in a vacuum manifold (cat: 9014579; Qiagen) and applied vacuum pressure at 800 mbar using a Millipore XX60-220-50 Vacuum Pressure Pump for 10 minutes. Membrane-bound DNA was resuspended in 55 µl of ultrapure water by vortexing the plate at low-speed (~600 rpm) on an appropriate multiplate adapter. We sequenced 2×25 µl of the DNA suspension, with sequencing primers oF and oR respectively. Sequencing was performed using a ABI 3730XL sequencer at the Genome Analysis Centre (TGAC, Norwich, UK).

DNA sequences were reverse-complemented when required and both strands were aligned to a reference sequence (Pasteur MLST website:

www.pasteur.fr/recherche/genopole/PF8/mlst/) using a web-based version of the MUSCLE algorithm www.ebi.ac.uk/Tools/msa/muscle/) and manually trimmed using the BIOEDIT software (Hall, 1999). All sequences aligned correctly, with no gaps. Novel alleles and sequence types (STs) were defined using the Pasteur database. When a new polymorphism was detected, we checked the ABI trace file for a clear signal, confirming the new SNP. When there was ambiguity in the sequencing signal, we reamplified and sequenced fragments. To include ECOR strains in our analyses, we retrieved 66 sets of 8 sequences already present in the Pasteur database.

2.4.3. Intraspecies diversity estimation using the EstimateS freeware

The main unit of diversity within *E. coli* isolates tested by MLST is the sequence type (ST) as illustrated in a recent study (de Muinck et al., 2011). When comparing strains from different environments, it is interesting to examine if there is variation in intra-population diversity, to infer on possible selection bottlenecks on diversity imposed by these different environments. A simple way of doing so is to calculate the rarefaction in the sample, as well as other diversity estimators, for which we used the EstimateS version 8.2.0 freeware (Colwell, R. K.; http://purl.oclc.org/estimates).

2.4.3.1. Diversity estimation

- **Chao richness estimator**

A non-parametric method to estimate diversity and richness was proposed by Anne Chao (1984). The classic Chao estimator is calculated as follows (Chao, 1984):

$$\widehat{S}_{Chao} = S_{obs} + \frac{a^2}{2b}$$

where S_{obs} is the number of observed species (or any taxonomic unit, in our case ST); *a* is the number of species observed just once (or singletons), and *b* is the number of species observed just twice (or doubletons).

For the classical calculation of the Chao estimate *a* and *b* are both strictly positive. This calculation is commonly refered to as the Chao-1 estimate of richness. Chao-2 refers to the application of Chao-1 to several collections rather than one. In that case, singletons and doubletons (representing for the Chao-1 estimator calculation the observation of distinct species just once or twice in the collection) correspond for Chao-2 to the observation of species in just one or two different collections. The formula remains unchanged. A sample-size bias correction was proposed as below, and is computed by default in EstimateS (and used in this work):

$$\widehat{\boldsymbol{S}}_{\boldsymbol{Chao}} = \boldsymbol{S}_{\boldsymbol{obs}} + \frac{\boldsymbol{a}(\boldsymbol{a}-\boldsymbol{1})}{\boldsymbol{2}(\boldsymbol{b}+\boldsymbol{1})}$$

Additionally, EstimateS calculates a 95% confidence interval on the Chao estimator (both bias-corrected and uncorrected versions) using the following:

$$\boldsymbol{Lower\ 95\%\ value} = \boldsymbol{S}_{\boldsymbol{obs}} + \frac{\widehat{\boldsymbol{S}}_{\boldsymbol{Chao}} - \boldsymbol{S}_{\boldsymbol{obs}}}{\boldsymbol{K}};$$

$$\boldsymbol{Upper\ 95\%\ value} = \boldsymbol{S}_{\boldsymbol{obs}} + \boldsymbol{K}(\widehat{\boldsymbol{S}}_{\boldsymbol{Chao}} - \boldsymbol{S}_{\boldsymbol{obs}});$$

where $K = exp\left[1.96\sqrt{ln\left(1+\frac{var(\hat{S}_{Chao})}{(\hat{S}_{Chao}-S_{obs})^2}\right)}\right]$.

- **Abundance-based coverage estimator and rarefaction curves**

The ACE estimator calculation employs more complicated equations to estimate diversity. The reader can refer to appropriate references for more details (Colwell, R. K.; http://purl.oclc.org/estimates). The ACE estimator has been found to be slightly more accurate than the Chao estimator, but the two are typically shown together in

diversity studies (de Muinck et al., 2011), and are calculated jointly using the EstimateS software. Similarly to the ACE estimator, the equations to calculate rarefaction curves are complex but the EstimateS freeware automates their calculation (Colwell et al., 2004).

2.4.4. ClonalFrame analysis and construction of phylogenetic trees

Most of the published basic phylogenetic analyses and phylogenetic tree representations rely on quick reconstruction methods, such as neighbour-joining (NJ) or maximum likelihood (ML). For rapid or ectopic analyses, these methods are often fine and easy to perform. However, their biggest pitfall is that they do not take into account the perturbation of the phylogenetic signal by homologous recombination. For instance, if two phylogenetically distant isolates share some alleles used in multilocus analyses because of some recent recombination events, methods such as NJ and ML, having very similar sequences as an input, will assume those two isolates are closely related, independently of their real genealogical ancestry. A Bayesian-based method implemented in the ClonalFrame software (Didelot and Falush, 2007), seeks to address this problem and produce clonal genealogies of a set of isolates, while identifying and minimising the "recombination noise" blurring the phylogenetic signal. ClonalFrame uses multilocus sequence data as input, or even a small number of full genomes (Didelot and Maiden, 2010). As recombination information is taken into account, ClonalFrame suggests phylogenetic relationships between strains based on the whole set of tested isolates provided. When isolates are removed or added from that set, the topology of the output trees can vary, as for instance ClonalFrame may

find or lose useful information to define or not a particular branch of the tree. It is important to bear in mind that in contrary to NJ or ML methods of phylogenetic reconstruction, ClonalFrame will not "force" relationships of closely related isolates. If two isolates in a ClonalFrame tree are found to be closely related, but information is missing to relate them further (possibly because the recombination signal is too high), they will simply be placed at equidistance of the same node. For these reasons, ClonalFrame is a much more appealing way to accurately represent phylogenetic relationships between same-species isolates and therefore is now commonly used as a way of producing trees based on MLST data (Jaureguy et al., 2008; Didelot et al., 2009; Didelot et al., 2011).

In this study, we used the ClonalFrame software to reconstruct the clonal genealogy of our strains. For each isolate, sequences were first concatenated in FASTA format and then formatted for input into the ClonalFrame software (Didelot and Falush, 2007). ClonalFrame was run with default settings and a Newick consensus tree was exported from the output. Some of the trees presented in this work were visualised and annotated in MEGA5 (Tamura et al., 2011), others online at the iTOL website (http://itol.embl.de/) on which trees imported in the Newick format can be easily annotated (Letunic and Bork, 2011) to produce publication-ready figures.

2.5. Microarrays-based comparative genomic hybridisation (CGH)

2.5.1. Principle of CGH

The use of miniaturised DNA microarrays was first reported in 1995 (Schena et al., 1995) and since then became one of the most affordable and practical methods to enable the comparison of genomes and transcriptomes. Microarrays physically consist of a high density print of DNA amplicon probes on a glass slide. One can then experimentally hybridize genomic DNA (gDNA) or complement DNA from messenger RNA to repectively examine and compare genomes and the transcriptomic state of cells in given conditions. In the work presented below, we used microarrays to compare the genomic content of multiple wild-type strains of *E. coli*.

Briefly, gDNA from a strain to test and DNA from a reference strain (typically a strain or a mixture of DNA encompassing all probes printed on the microarray used) are labelled with two different fluorescent dyes, mixed and jointly hybridised on the microarray. After scanning of the microarray, spots showing a mixture of both fluorescent labels correspond to genes present in the tested strain, and spots showing only single fluorescent labels correspond to genes present in only the sample strain, or the reference. In order to carry out good and homogenous comparison between strains, one should minimize the presence of genes that are present in the tested strain but not in the reference, but in the case of multi-genome or custom-made microarrays this is sometimes impossible. This was the case in our study using ShEcoliO157 microarrays, for which an additional signal correction procedure was developed.
Using CGH, we analysed the hybridised genomes of 20 ECOR strains (ECOR-02, 04, 07, 10, 17, 23, 24, 28, 30, 34, 38, 41, 45, 55, 58, 62, 67, 68, 71 and 72) and 21 GMB

(GMB02, 07, 16, 18, 23, 34, 41, 46, 54, 58, 59, 61, 64, 66, 74, 78, 81, 88, 91, 92, 100).

2.5.2. ShEcoliO157 microarrays

2.5.2.1. Description of microarrays and printing procedure

ShEcoliO157 multi-genomic microarrays are an upgraded version of previously published microarrays, and were designed at IFR (Anjum et al., 2003; Lucchini et al., 2005). Each microarray comprise 6379 amplicons probes encompassing the complete genomes of *E. coli* K-12 strain MG1655 (4265 specific "b" genes), *E. coli* O157:H7 strain EDL933 (1128 specific "Z" genes), *Shigella flexneri* 2a strain 301 (555 specific "SF" genes) and a selection of various *E. coli* virulence genes (431 "VIRECO" genes). Microarrays were printed using the IFR Microarray Facility Stanford-type microarrayer (Thompson et al., 2001) by Yvette Wormstone and Carl Harrington in 2006 and 2009, respectively.

2.5.2.2. Reference strain used

Because of the composite nature of the ShEcoliO157 microarray, it was not possible to use a reference that would cover all the probes. As ShEcoliO157 microarrays harbour various known *E. coli* virulence genes, the use of a strain of the pathogenic serovar O157:H7 could represent a theoretical close match. We used a *E. coli* O157:H7 strain Sakai deletion mutant in *stx1,2* genes (constructed and kindly given

by Dr Martin Goldberg, University of Birmingham, UK) in order to allow manipulation with the appropriate safety in our cat-2 laboratory.

2.5.3. Protocol used in this study

2.5.3.1. gDNA extraction

Because gDNA preparations purified using spin column-based kits generate poor quality hybridisations on microarrays in our hands, we extracted genomic DNA using the gravity column-based Qiagen Genomic-tip DNA Extraction kit (cat: 10223; Qiagen). Additionally, we observed that following the manufacturer's instructions often did not completely lyse cells, leading to the clogging of the gravity columns, and thus to a waste of time and material. To optimize cell lysis for genomic DNA extraction, bacterial pellets were frozen at -80°C for at least one night. The next day, frozen cell pellets were thawed in the B1 lysis buffer (containing 100 mg/ml RNase A) from the kit. Extractions were then performed according to the manufacturer's instructions with the exception of longer lysis incubation times (3-4 hours instead of 30 min for the first lysis step at 37°C and 2-3 hours for the second lysis step at 50°C). DNA concentration was measured using a Nanodrop (Labtech Ltd, Ringmer, UK). Microarray blocking, gDNA labelling, hybridisations and washes were performed according to the protocol developed by the IFR *Salmonella* group (http://www.ifr.ac.uk/safety/microarrays/) and are detailed below.

2.5.3.2. Microarray blocking

Microarrays were printed on epoxysilane-treated glass slides (Corning GAPS II; cat: CLS40005; Sigma-Aldrich) to allow direct, correct and precise printing of DNA onto them (Taylor2003). Therefore, the slides need to be "blocked" prior to utilisation, to prevent unspecific binding of labelled DNA, i.e. elsewhere than on the DNA probes. The principle of any blocking procedure is to alter the epoxysilane coating to reduce nonspecific interaction with DNA. In this work and more generally in the Salmonella group at IFR, the blocking reagent used is 1,2-dichloroethane (DCE; cat: 284505; Sigma-Aldrich).

The two microarrays printed on each slide are usually visible to the naked eye but become invisible after blocking as the salts present in the printing solution dissolve during the blocking procedure. The position of the two microarrays present on each slide was therefore marked with a diamond-tipped pencil prior to DNA immobilisation. In order to strongly immobilize DNA probes on the microarray, slides were irradiated twice using the "auto cross-link" setting (corresponding to a 2×120,000 microjoules/cm2 irradiation) in a UV cross-linker (Stratalinker UV Crosslinker, Stratagene). Slides were then incubated for 1 hour at room temperature with gentle agitation in a "blocking solution" composed of 0.5% (w/v) succinic anhydride (cat: 239690; Sigma-Aldrich) in 300ml of anhydrous DCE containing 3.75ml of 1-methylimidazole (cat: M50834; Sigma-Aldrich). Slides were then transferred for 2–3 minutes in fresh DCE, 2 minutes in boiling water and 1 minute in 96% ethanol before being dried by centrifugation at 1,200 rpm for 5 min.

2.5.3.3. gDNA labelling

DNA labelling is performed using the Klenow fragment to incorporate fluorescently tagged deoxyribonucleotides in the 3' to 5' direction from random primers annealed along the target DNA. In this work, we used 1 µg of each DNA species per microarray hybridisation (2 µg total DNA).

Purified gDNA was vortexed for a few seconds to fragment the DNA. In a sterile tube, 1 µg of DNA (test or reference) was vacuum-concentrated and resuspended in 10 µl of sterile molecular biology grade water (cat: W4502; Sigma-Aldrich) and 10 µl of 2.5X Random primer/reaction buffer mix from the BioPrime DNA Labelling kit (cat: 18094-011, Invitrogen) were added. Samples were boiled for 5 min to denaturate DNA and put on ice for 5 min to allow reannealing of the random primers. Then, still on ice, we added 2.5 µl of 10X dNTP mix (1.2 mM of dATP, dTTP, dGTP; 0.6 mM of dCTP; 10mM Tris pH8.0; 1mM EDTA). We also added 1.5 µl of 1mM of the corresponding fluorescent label (usually Cy5-dCTP for reference DNA and Cy3-dCTP for test DNA; cat: PA55321, GE Healthcare Lifesciences) and 0.5 µl of the Klenow fragment from the BioPrime kit. The total reaction volume is 24.5 µl. The reaction mixtures were incubated at 37°C overnight protected from light. Labelled DNA was purified using a QIAquick PCR Purification Kit (cat: 28104; Qiagen) to remove unincorporated fluorescent Cy-dyes, with the precaution of eluting twice using 2×50 µl of sterile water to maximise recovery from the membrane of the spin-column. Samples were vacuum-concentrated and resuspended in 9.75 µl of sterile water. Test and reference DNA were combined in a 1:1 mix.

2.5.3.4. Microarray hybridisation and washes

To the 9.75 µl mix of labelled test and reference DNA, we added 1.125 µl of 25 mg/ml yeast tRNA (cat: R8759; Sigma-Aldrich), 2.25 µl of 20X saline-sodium citrate (SSC) buffer, 0.36 µl of 1M 4-(2-hydroxyethyl)-1-piperazine-ethane-sulfonic acid (HEPES) pH 7.0 buffer, 1.5 µl of Denhardt solution and 0.338 µl of 10% sodium dodecyl sulfate (SDS). The role of tRNA with Denhardt solution and SDS, both mixtures of high-molecular weight polymers, is to increase specific binding of labelled DNA on the probes by saturating non-specific binding sites. Hybridisation mixes are incubated at 99°C for 2 min and left at room temperature for 5 to 10 min. This step denatures labelled DNA into single stranded molecules. Tubes are then centrifuged 2×10 min at maximum speed to pellet precipitated SDS. Slides are placed in metal hybridisation chambers and hybridisation reactions are transferred to the microarrays (the final volume per array is 15.32 µl). A coverslip is added and 30 µl of SSC buffer is added to maintain adequate humidity in the chamber. Hybridisation occurs for 14 to 18 h (typically overnight) at 63°C in a water bath. After incubation, and quickly disassembling hybridisation chambers, slides are washed in 2x SSC buffer containing 0.1% SDS at 68°C for 5 min, followed by 1x SSC at room temperature for 5 min twice on an orbital shaker at 60 rpm, and finally in 0.2x SSC at room temperature for 5 min twice on an orbital shaker at 60 rpm. Slides are then dried by centrifugation at 1500 rpm for 5 minutes, and scanned within 3 hours using a GenePix 4000B microarray scanner (Molecular Devices, Inc.). Grid alignment and signal normalisation are performed using BlueFuse for Microarrays v3.6 (BlueGnome, Cambridge, UK). In this work, a second hybridisation was realised for each tested *E. coli* isolate, as a technical replication.

2.5.3.5. Raw signal correction

One would imagine that when gDNA is hybridised on a microarray, there are only 2 kinds of results, either there is a fluorescent signal at a given spot and the corresponding gene is considered present, or signal is missing and the corresponding gene is considered absent. In practice, distribution of signals are different, with a high peak of high intensity signals, corresponding to present genes, with a tailing distribution of lower intensity signals of genes that are absent or too divergent to anneal correctly to their corresponding probes. Additionally, for the analysis of microarray results, only the log-ratio of fluorescence signals between the reference DNA channel and the test channel for each gene matters. Using an absolute reference DNA encompassing all probes on the microarray, it is then easy to identify genes that are present in the reference but not in the tested isolate (**Figure 2.2**).

In our case however, the reference DNA was a gDNA extraction from cultures of *E. coli* O157:H7 Δ*stx1,2* which is not encompassing all probes present on a ShEcoliO157 microarray. This non-absolute coverage of the probes by the reference comes with the possibility that a gene is present in the test isolate but absent from the reference, something which is not happening with absolute references. Therefore, it is not possible to use ratios to discriminate between genes present and absent from the tested both isolate and the reference:

$$log_2\left(\frac{Present\ in\ reference}{Absent\ in\ test}\right) > 0;\ log_2\left(\frac{Absent\ in\ reference}{Present\ in\ test}\right) < 0;$$

$$log_2\left(\frac{Present\ in\ reference}{Present\ in\ test}\right) = log_2\left(\frac{Absent\ in\ reference}{Absent\ in\ test}\right) = 0$$

To overcome this, we developed a signal correction procedure, of which we provide here a step-by-step description. First, we gathered fluorescent signals from 92 independent *E. coli* O157:H7 strain Sakai gDNA hybridisations and performed a basic normalisation for all of them by dividing each signal value by the sum of all signals of the same hybridisation. We obtained the following frequency plot shown in **Figure 2.2**.

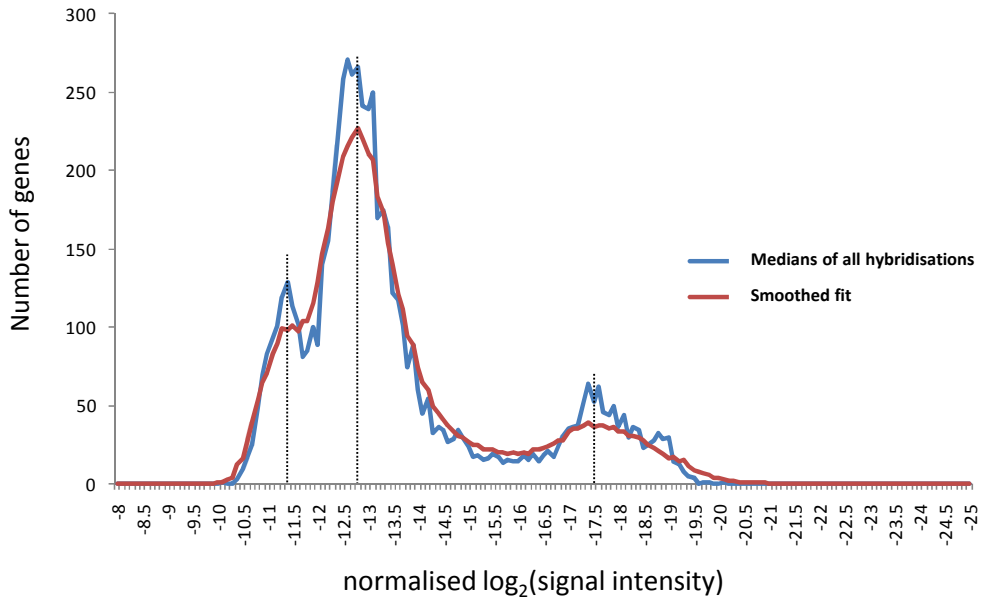


Figure 2.2. Frequency plot of ShEcoliO157 microarray hybridisation signals and its smooth fit ("medians of medians").

Frequencies in signal intensities peaked according to 3 different populations: present genes (high frequency peak "1"), absent or highly divergent genes (smaller frequency peak "2"), and a "shoulder" peak of present genes, "3"). We observed that this "shoulder" of present genes was composed almost exclusively by Z genes, belonging to the *E. coli* O157:H7 EDL933 genome subset of ShEcoliO157 microarrays (**Figure 2.3**).

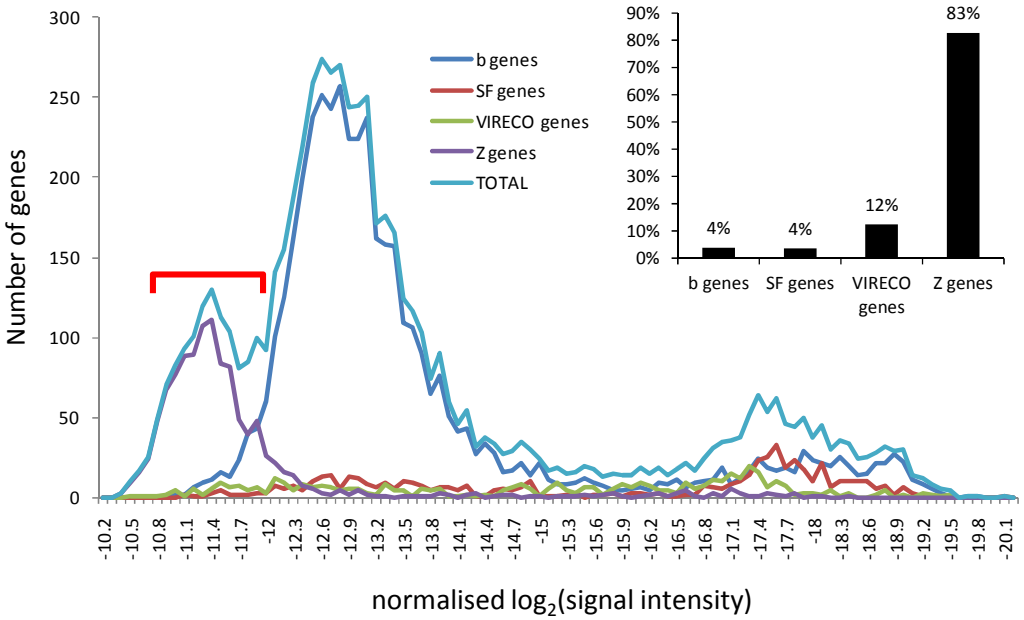


Figure 2.3. Frequency plot of microarray hybridisation signals for different groups of genes. The inset represents the proportion (%) of gene groups in the "shoulder" of the biggest peak (see text).

The inset in the plot corresponds to the proportion of total genes from the different ShEcoliO157 genome subsets present in the shoulder (indicated in red on the previous frequency plot between intensities -10.8 and -12). More than 80% of all Z genes on the microarray had a higher than average signal intensity and located in the "shoulder". This probably reflects differences in the probe design, with O157 probes providing higher signals.

The normalised signal intensities were centred by offsetting the "present" peak so that it would be centred on 1. The "shoulder effect" was also attenuated by offsetting with a lower value, so that most of the present genes would be in the higher "present" peak. On the frequency plots below, the frequency of normalised signals before correction

and centring is representing in blue and the frequency of normalised signals after centring and attenuation of the "shoulder effect" is represented in red (**Figure 2.4**).

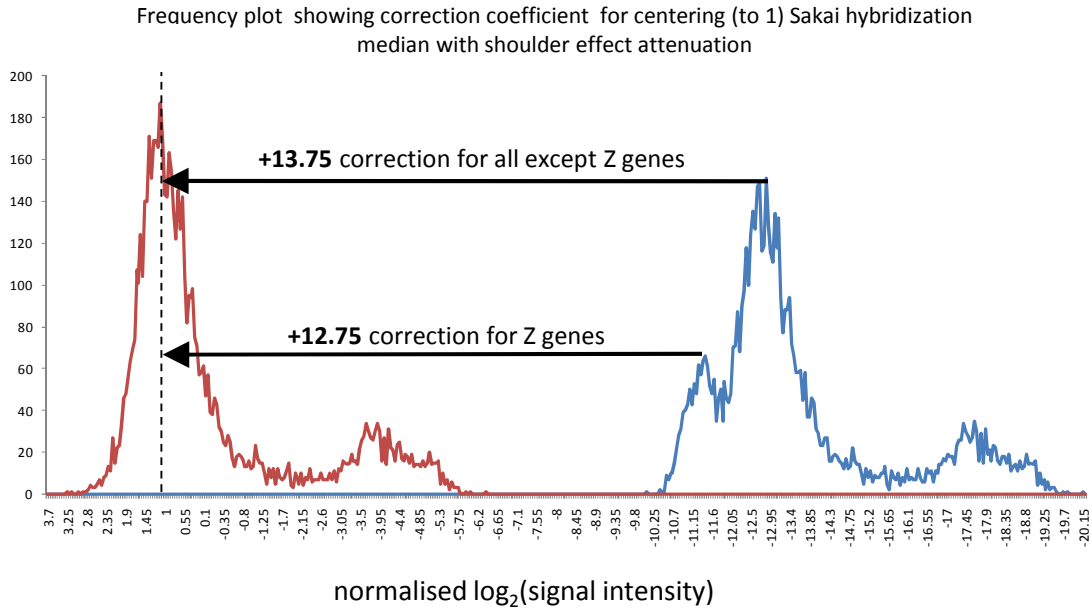


Figure 2.4. Representation of the offset modification to center signal values and attenuate the "shoulder effect" during correction of hybridisation signals (see text). The blue curve represents the frequency plot of hybridisation signals before correction, and the red curve is after correction.

With the basic normalisation and centring steps, the problem of ratios is corrected, and we can import directly this dataset in any subsequent analysis software. Present genes have now a normalised value for signal intensity of roughly 1, and the normalised value for signal intensity of absent or divergent genes is now negative.

We applied the same correction steps presented above to all other individual hybridisations performed in this study and obtained the following frequency plot, showing an overall uniformity of signal frequencies and a relative normalisation of the "shoulder effect" (**Figure 2.5**).

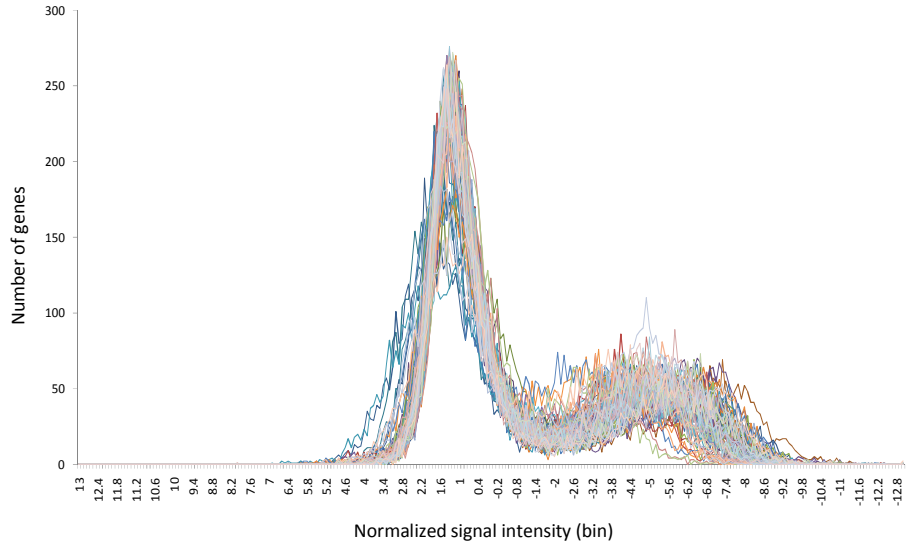


Figure 2.5. Corrected combined frequency plots of hybridisation signals for all hybridised strains of *E. coli* tested in this study.

2.5.4. Dynamic cut-off for gene presence

The ideal output for gene content comparison and analysis is binary, with genes that are present in the tested strain, and genes that are absent or too divergent to hybridise correctly on the microarray. This requires the determination of a threshold value (or "cut-off") from signal distributions to determine gene presence. In early CGH studies, this threshold value was constant and determined empirically without taking into account the variation of signal intensities distribution between isolates. But because distributions do slightly vary between samples and replicates, a threshold valid for one distribution may produce false negatives or positives when applied to another. In this study, we used the "Genomotyping Analysis by Charlie Kim" (GACK) software (Kim et al., 2002) to define dynamic cut-off values for each distribution and create a binary dataset accordingly.

Briefly, we averaged both technical replicate signals for each tested strain and formatted an input file according to the GACK manual and a recent comprehensive review of this process (Stabler et al., 2010). In GACK, the dynamic cut-off determination is dependent of the "estimated probability of presence" (EPP) value, calculated as follows:

$$EPP(\%) = 100 \times \frac{Predicted\ curve\ for\ complete\ presence}{Observed\ curve\ of\ signal\ intensities}$$

The percentage of EPP (below in blue) is defined by dividing a predicted curve for which all genes are present ("EPP curve" different for all samples, below in black) by the observed data distribution (in red in **Figure 2.6** with the example of observed frequencies of signal intensities for strain GMB106):

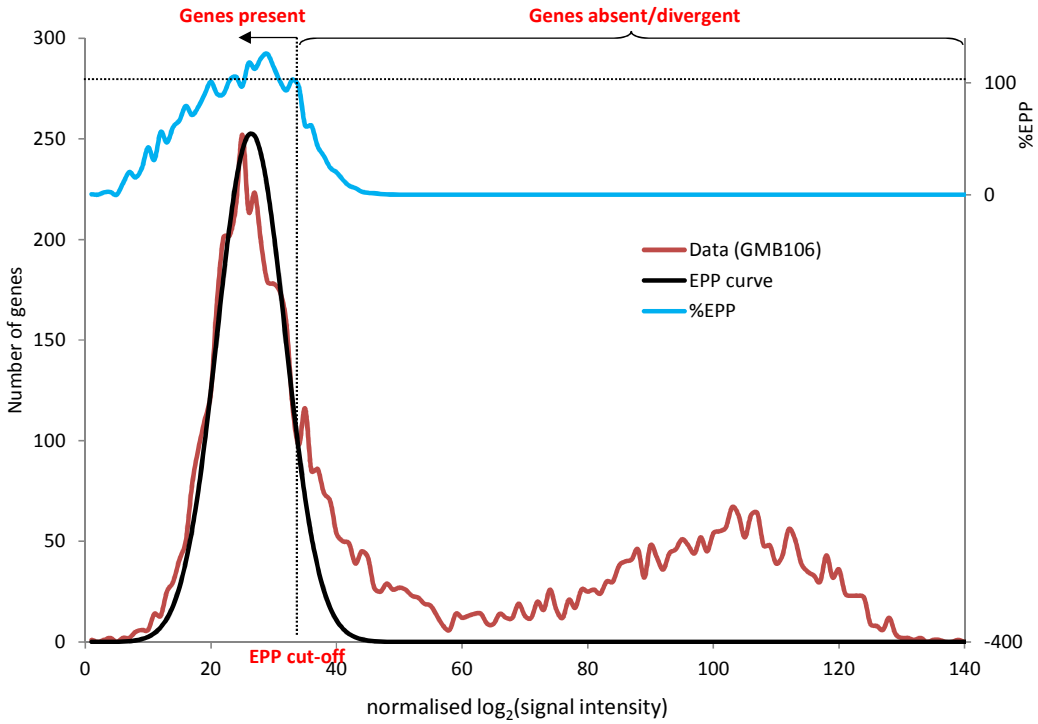


Figure 2.6. Graphical illustration of GACK dynamic cut-off procedure (see text).

A particularity of GACK is that as soon as the %EPP curve reaches 100% (blue curve above, starting from the right), its shape becomes variable and unpredictable so only the portion of the curve on the right is considered valid for analysis.

Users of GACK can define their desired value for %EPP, therefore for the dynamic cut-off determination. When the %EPP curve reaches the value defined by the user, it then defines the cut-off value, from which all genes on the left are considered present and all genes on the right are considered absent or divergent. A value of EPP=100%, as illustrated above, will set a cut-off value close to the peak of present genes (for GMB106, around 35) and allow for an analysis limiting the number of falsely positive genes. Conversely, if EPP is set to 0%, the cut-off value will be set further away from the peak of present genes, and will allow an analysis limiting the number of falsely negative genes. The estimated presence of genes between EPP=0% and EPP=100% is unknown, and a setting of GACK (called “trinary output”) allows to assign them a particular value for information purposes. For our analysis, we decided to minimise false positives and therefore used the EPP=100% setting in GACK, with a binary output.

2.5.5. Genetic association tests with various parameters

In order to identify genes across our tested strain set whose pattern of presence or absence after the GACK process correlated to various parameters, such as phylogenetic groups, the source or location of isolation or metabolic profiles on Biolog plates, we used non-parametric association tests (Mann-Whitney-Wilcoxon

when comparing 2 groups, and Kruskal-Wallis when comparing more than 2 groups). When attempting to associate patterns of genes with normally distributed phenotypes (for instance, siderophore production), we used a parametric test (Student's t-test). These statistical tools are implemented in Genespring v.7.3 and were used that way.

We kept a high stringency filtering of the results to limit false positives for the association analysis by applying multiple testing corrections (MTC) to the obtained *p*-values list. Indeed, because of the nature of statistical tests, if 6500 genes are examined at a *p*-value cut-off of 0.05 and without MTC, the expected level of genes identified as significantly positive by chance only is:

$$Expected\ false\ positives = p\ \times N = 0.05\ \times 6500 = 325\ genes$$

When mentioned in the text, we used the Benjamini-Hochberg MTC on test results, as recommended by the Genespring manual, with a threshold at *p*=0.01 (Benjamini and Hochberg, 1995). Briefly, after a test is computed, the *p*-values of each gene are ranked from the smallest to largest and the following nth *p*-values are corrected as follows (the 1st p-value remains unchanged):

$$p_{corrected} = p_{uncorrected}\ \times \frac{n}{n-1}; for\ n > 1$$

In our case, for each $p_{corrected} < 0.01$, the corresponding genes were considered significantly associated with the tested parameter.

2.6. Carbon metabolic profiling using Biolog plates

2.6.1. Principle of the Biolog system

Biolog phenotype arrays were developed to allow the simultaneous comparison of multiple metabolic phenotypes per single culturable isolate (Bochner, 2009). Individual tests in 96-well pre-prepared plates consist of a mixture of a compound (carbon, nitrogen, phosphorus, sulphur source or else) combined with a redox chromogenic dye, triphenyl tetrazolium chloride (TTC) and a Biolog proprietary incubation medium. For each test, the compound to be tested is present as a sole metabolic source of its kind. For instance, if carbon metabolism is to be tested, all elements required for growth excluding carbon will be provided. In aerobiosis, when a metabolic compound is taken up and transported in a cell, respiration will initiate and nicotinamide adenine dinucleotide (NADH) is generated, which in turn generates reducing power in the cell (Bochner, 2009). TTC turns blue/violet upon direct reduction by NADH. This reaction is irreversible (Bochner and Savageau, 1977) and TTC reduction is used as a colorimetric indicator for the ability to respire on the various compounds present in a Biolog plate. The applications of the Biolog system are multiple and can range from the comparison of metabolism between strains of the same species (Sabarly et al., 2011) to microbial identification (Pinot et al., 2011) and community metabolic profiling (Bossio and Scow, 1995; Di Giovanni et al., 1999). In this study, we used Biolog GN2 microplates (Techno-Path, UK) encompassing 95 C-sources representative of the requirements to discriminate and identify most Gram-negative bacteria.

2.6.2. Biolog data analysis and statistics

As the Biolog analysis procedure we developed in this work (in collaboration with E. Kate Kemsley) represents a novel approach, we included details of the analysis (including threshold determination) in the results section. For statistical calculations, we used MATCAD with the Statistics module installed, and GraphPad Prism version 5.01. Representation of graphs was performed using Microsoft Excel 2007/2010 or GraphPad Prism version 5.01.

2.7. Phenotypic analyses of colonisation-associated traits

2.7.1. Biofilm formation on polystyrene surfaces

Biofilm formation was assessed using crystal violet staining of bacteria attached to the polystyrene surface of a 96-well microtitre plate. Bacteria were grown in LB overnight at 37°C. The next day, optical density was standardised to OD_{600}=0.175, and this suspension was diluted 1:10 in 100 µl of colony-forming-antigen (CFA) medium (Evans et al., 1977). Static cultures were grown in 96-well microtitre plates at different temperatures for up to 72 h (biofilm formation at 28°C was monitored after 72 h, at 37°C after 48 h).

At the desired time point, plates were washed 2 times using sterile water, blotted on absorbent paper after each wash to remove excess water. Plates were dried at 60°C for 30 min, or until complete dryness was observed. Plates were filled with 130µl/well of a 1% filtered crystal violet solution, and incubated for 30 min at room temperature.

Plates were emptied, washed 3 times using sterile water, blotted on absorbent paper after each wash, and then dried for 1h at 37°C. Crystal violet was dissolved by adding 150 µl/well of a 20:80% acetone/ethanol mix and left at room temperature for 10 min. Optical density was measured at 590-600nm.

2.7.2. Bacterial motility

We were interested in assessing the proportion of motile and non-motile strains in our different *E. coli* collections. The ability of bacteria to swim is a phenotype easily assessable using low concentration agar gels. We used commercially available soft (0.4% w/v) agar tubes (BBL™ Motility Test Medium, Becton Dickinson) to detect bacterial motility on pure culture of *E. coli* isolates. Fresh colonies on LB plates incubated overnight at 37ºC were picked at their centre with a flame-sterilised platinum wire, which was then used to stab BBL™ Motility Test Medium tubes. Motility tubes were then incubated at 37ºC for 24 h before visual inspection of the motility phenotype. When the tested strain was non-motile in the tested condition, bacterial growth was visible inside the stab mark, but not elsewhere in the tube. Conversely, when the tested strain was motile, growth extended in the soft agar, away from the stab mark. Motility was visually scored, and a second biological replicate was performed.

2.7.3. Siderophore production

The ability to produce and secrete iron-scavenging molecules called siderophores is known to occur in many environmental organisms for which iron is a limiting factor

of growth. *E. coli* is known to produce up to 4 siderophores: enterochelin (or enterobactin), salmochelin, yersiniobactin and aerobactin, respectively encoded by the *ent*, *iro*, *irp* and *iuc* operons (Valdebenito et al., 2006). We used chrome azurol S (CAS)-based solid medium to assess siderophore production *in vitro*. CAS is an indicator molecule with moderate affinity for iron molecules. When CAS is bound to iron, the complex is dark blue to green, the presence of siderophore (for which iron has a greater affinity than for CAS) disrupts the CAS binding with iron, which then turns bright yellow. In our assay, we grew colonies of strains on CAS indicator agar for 48h at 37°C and measured the halo diameter as a linear proportion of the ability to produce siderophore using the ImageJ software.

The recipe we used for CAS indicator agar preparation is derived from a previously published protocol (Payne, 1994). This complex medium relies on the mixture of 3 independently prepared solutions: a 10X modified M9 (MM9) salts solution, a CAS-HDTMA solution (which is a mix of 2 other solutions) and a deferrated casamino acids solution. To prepare the 10X MM9 solution, 3 g of KH_2PO_4, 5 g of NaCl and 10 g of NH_4Cl are added to 1 l of milliQ water and autoclaved. The CAS-HDTMA solution is prepared by first dissolving 605 mg of CAS powder in 500 ml of milliQ water, to which are added 100 ml of a ferric solution composed of 1 mM FeCl3 in 10 mM HCl; and then this mixture is added to 400 ml of a HDTMA buffer solution (729 mg of HDTMA powder in 400 ml of milliQ water heated at 40°C) and autoclaved. The deferrated casamino acid solution is prepared by first dissolving 10 g of casamino acids in 100 ml milliQ water. Casamino acids are then extracted from this solution by adding 100 ml (equal volume) of 3%(w/v) 8-hydroxyquinoline in chloroform to remove contaminating iron, and then extracted again with 100 ml (equal volume) of

chloroform to remove traces of 8-hydroxyquinoline. Finally, CAS indicator agar is prepared by first adding 6g of NaOH (dissolved in the water), 30.24 g of PIPES buffer, 100 ml of 10X MM9 salts solution and 15 g of agar to 750 ml of milliQ water to be autoclaved and kept at 50°C. When the agar is still liquid, 30 g of deferrated casamino acids solution are quickly added along with 10 ml of filtered-sterilised solution of 20% glucose, 1ml of 1M $MgCl_2$ and 1ml of 100 mM $CaCl_2$. The solution is mixed, 100 ml of CAS-HDTMA solution are added and plates are poured and dried before utilisation. Colonies of siderophore-producing bacteria grown on this medium are surrounded by a yellow or orange halo (**Figure 2.7**).

Figure 2.7. Example of halo obtained by a siderophore producing strain streaked on CAS indicator agar. The strain used is *E. coli* strain GMB30.

2.7.4. Multiplex PCR for detecting siderophore biosynthesis genes in *E. coli*

To develop a method to detect the 4 described siderophore operons, we used *E. coli* strain Nissle 1917, a probiotic strain commercialised under the brand name "Mutaflor" and kindly provided by Ardeypharm GmbH, which possesses the 4 operons in its genome (Valdebenito et al., 2006). We used the MPprimer website

(Shen et al., 2010) and sequences of enterobactin from *E. coli* K-12 strain MG1655 (gb: U00096.2), salmochelin from *E. coli* strain Nissle 1917 (gb: AJ586887.1), yersiniabactin from *Yersinia pestis* (gb: CP001585.1) and aerobactin from an *E. coli* plasmid sequence (gb:144692 and X76100.1) to design multiplex primers.

MPprimer found primer sequences for all tested genes, and produced the corresponding virtual electrophoresis patterns (**Figure 2.8**).

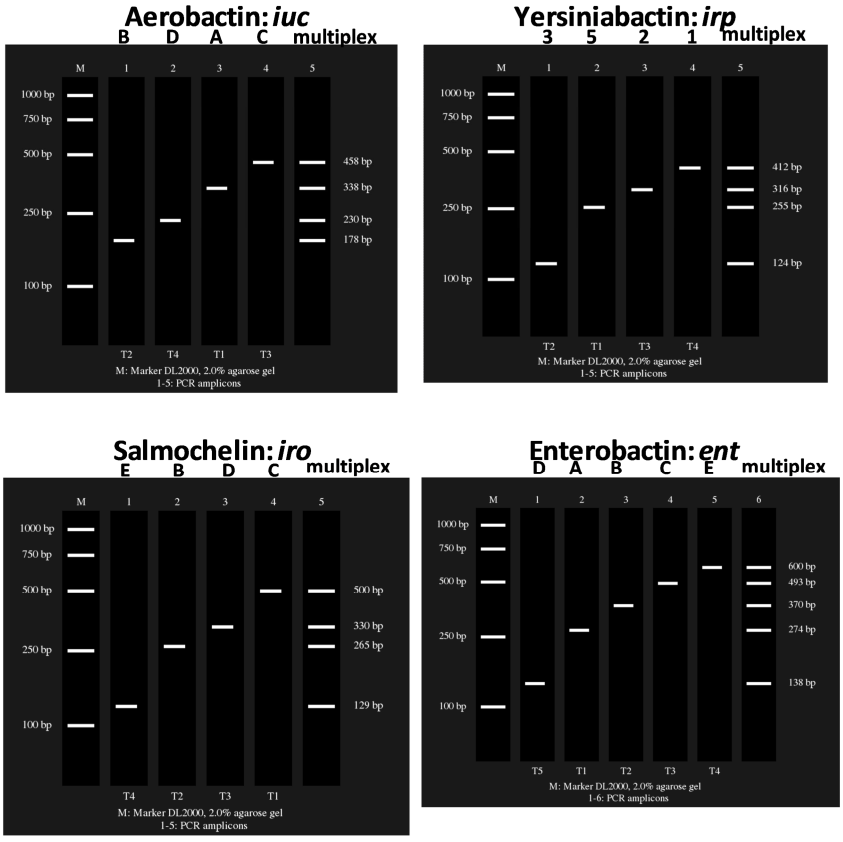


Figure 2.8. Virtual electrophoretic patterns for siderophore biosynthesis gene multiplex PCR calculated by the MPprimer website. Aerobactin profiles: *iucB, iucD, iucA, iucC*; yersiniabactin profiles: *irp3, irp5, irp2, irp1*; salmochelin profiles: *iroE, iroB, iroD, iroC*; enterobactin profiles: *entD, entA, entB, entC, entE*.

We ordered the obtained primers (**Table 2.5**).

Table 2.5. Primer sequences for siderophore biosynthesis gene multiplex amplification. The primers were ordered from Sigma-Genosys (UK).

	Primer target	Amplicon size (bp)	Tm (FP/RP) (°C)	Forward primer (FP) / Reverse primer (RP)
Aerobactin	*iucA*	338	59.9/60.0	5'-GGACTGCGCGATCTTAACGGCA-3' / 5'-AATCCGACCGGCAGATCCCCAA-3'
	iucB	178	60.1/60.0	5'-ATCTTGGCTGGGGGCTGGACAA-3' / 5'-ACCTGTCCGAAAAGCGTCAGCG-3'
	iucC	458	59.9/59.9	5'-AACGTGGTATCTGGGGCTGGCT-3' / 5'-ACTGAAGCGGGCAAACTCCTGC-3'
	iucD	230	59.6/59.7	5'-CGGATGGCATCACTGCCCGATT-3' / 5'-ACGCAGAACGGTAACCTGTGGC-3'
Yersiniabactin	*irp5*	255	60.0/60.0	5'-TATGGCGGAAGGCCTGCTCTGT-3' / 5'-CTGACATTGTCGCCTGTGCGGT-3'
	irp3	124	60.0/59.9	5'-TTGCTGGCTCTGGGCGTTGATG-3' / 5'-GATAGCGCTGAAGCAGCAGGCA-3'
	irp2	316	60.0/60.0	5'-ACTTCCTCGCCCGGCGTAATCT-3' / 5'-GCCGCAATGTGTGGCTGAGAGT-3'
	irp1	412	60.0/60.0	5'-CCATATTGGCCGCACGCTCGAT-3' / 5'-GGCGTTAAACCGTTCGGGCTGA-3'
Salmochelin	*iroC*	500	60.1/60.0	5'-GTCGGGCAATCACCAGCAGCAT-3' / 5'-GGGCGCATCGGGTTCAGGAAAA-3'
	iroB	265	60.1/60.1	5'-ATGGCCGAAGCCTACGGTTTGC-3' / 5'-CGGGTTGGTGGTGTTTGACGCT-3'
	iroD	330	60.1/60.0	5'-AGTGGCTGAGCACCAGACCGAA-3' / 5'-TCGTGCTGCCCGATGGTGAAAC-3'
	iroE	129	60.0/59.9	5'-TGCCTGGCGAAAGGAGGCATTG-3' / 5'-GCTGTCGGGGTGCTGTCGAAAA-3'
Enterobactin	*entA*	274	60.2/60.1	5'-GTCGGCGGTGCGTTTAACCTGT-3' / 5'-GCTGTTCTTCGGCGTCATCGCT-3'
	entB	370	59.9/59.8	5'-TGCGCGACTACTGCAAACAGCA-3' / 5'-TGCTCGTCACGGCTGAAATCGG-3'
	entC	493	59.9/60.1	5'-GAATGTGGTGGAACGCCAGGCA-3' / 5'-AGTTGCGAGATGCCACAGCGTC-3'
	entE	600	59.9/60.0	5'-ACGTCTTTCTGCCACCCTTGCG-3' / 5'-ACACACTCCACGCGATCCGGTA-3'
	entD	138	59.7/59.9	5'-ATTTAGCCGGACGGATCGCTGC-3' / 5'-ATGCCGTAGTCCCACAGTGGCT-3'

We tested different conditions of multiplex PCR on an *E. coli* strain Nissle 1917 genomic DNA preparation until we obtained satisfactory amplification of most of the bands **(Figure 2.9**).

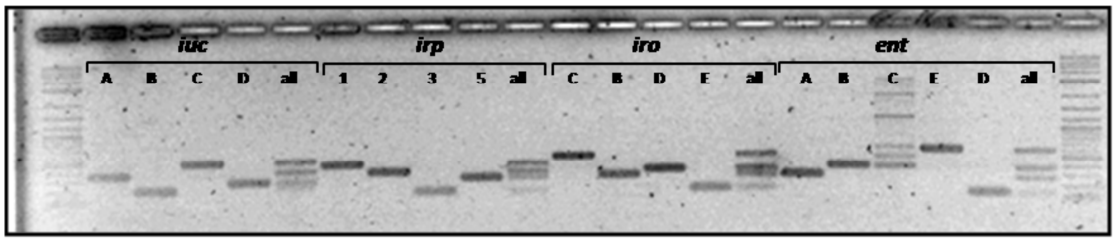


Figure 2.9. Electrophoresis gel after simplex and multiplex amplification of siderophore biosynthetic operons from *E. coli* strain Nissle 1917. Annealing temperature for the PCR was 50°C. DNA ladder is the 2-log DNA ladder from NEB.

Our selected amplification conditions were 1 cycle at 95°C for 5 min, 30 cycles at 95°C for 30s, 50°C for 30s and 72°C for 90s and 1 cycle at 72°C for 5 min. We used the GoTaq Green Mastermix (Promega) to prepare our amplification reaction. We could not amplify a single band for *entC* using these multiplex amplification conditions (**Figure 2.10**). However, we could amplify 4 genes out of 5 from the *ent* operon, which we judged enough for our detection purposes (**Figure 2.10**). Similarly and despite our prolonged efforts, *iucB* could be amplified in simplex but not in multiplex (**Figure 2.9**) so we excluded the corresponding primer pairs from the multiplex mix.

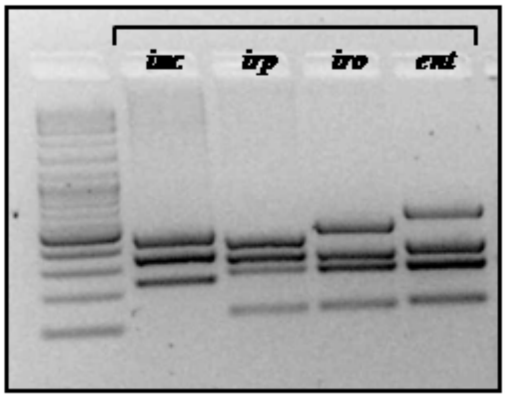


Figure 2.10. Electrophoresis gel pattern for a successful 15-gene multiplex detection of siderophore biosynthetic operons in *E. coli*.

In the results chapter, we present an application of this method to detect siderophore genes in 28 ECOR and GMB *E. coli* strains.

2.8. DGGE profiling of phyllosphere-associated bacterial community

2.8.1. Principle of community profiling

Bacteria populate agricultural environments in considerable amounts, reaching concentrations of 10^4-10^9 CFU/g in soil (Garbeva et al., 2004) and around 10^8 CFU/g on the aerial parts of plants (Lindow and Brandl, 2003). DNA extracted from the communities present in various ecosystems is usually used as a template for the PCR amplification of phylogenetic markers that can then be analysed by various methods including DGGE or HT-sequencing. In this work, we used DGGE to first examine microbial community profiles in agricultural soils and on various crops, and then to observe plant microbial community response to an artificial leaf contamination with *E. coli*. In this section, we describe the procedures used for sampling, DNA extraction and DGGE of epiphytic bacterial samples.

2.8.2. Environmental sampling

2.8.2.1. Field collection of soil and crops

The fields we sampled, located in East Norfolk (UK; **Figure 2.11**), were used for growing different crops, mainly spinach and rocket salad but also occasionally red

Batavia lettuce or red chard. We used ethanol-sterilised scalpels and hand trowel to sample leaf and soil material, which was placed in sterile stomacher bags on site. All samples were placed in a cooled ice box for transport back to our laboratory, where they were processed on the same day as presented in section 2.8.3.

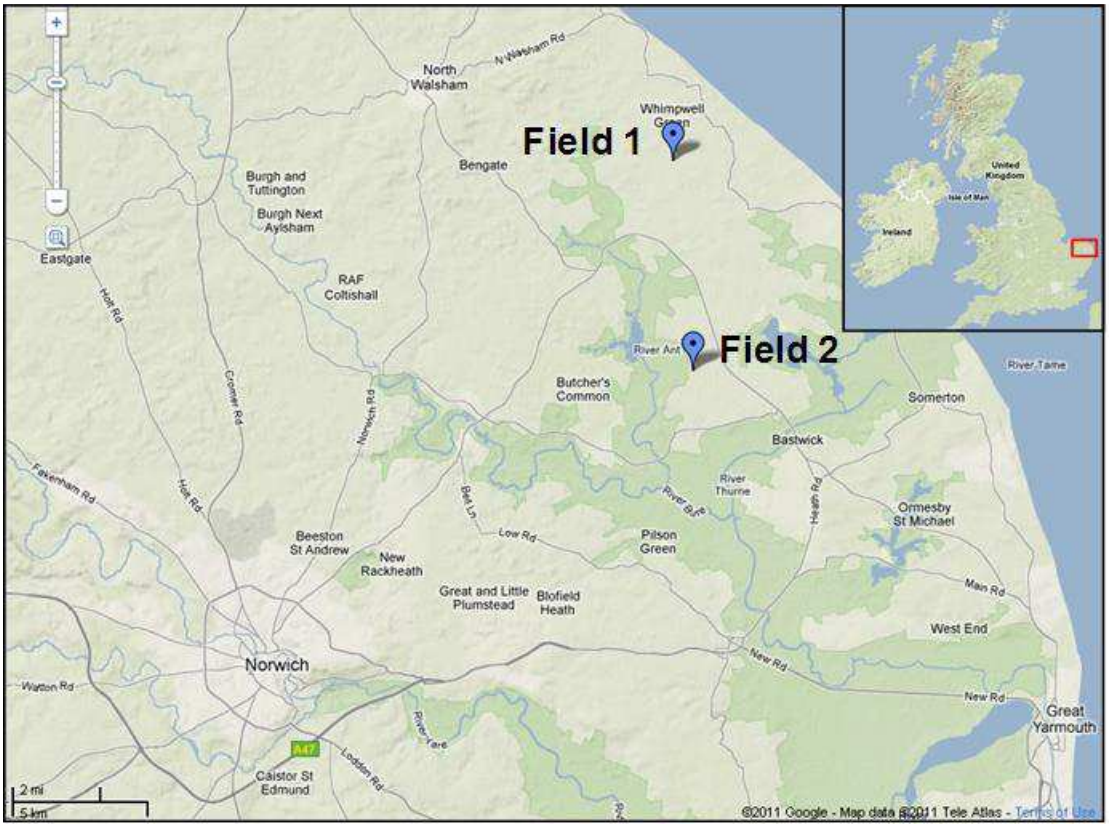


Figure 2.11. Geographical location of the agricultural sampling fields mentioned in this study. Fields are represented with a blue marker. This map is copyrighted by Google.

2.8.2.2. Chemical characterisation of tested field soils

We determined the textural and chemical characteristics of two field soils we sampled for microbial communities analyses. The British Standard BS3882:2007 procedure for soil analysis ("Specification for topsoil and requirements for use") was used by NRM Laboratories (Berkshire, UK) to analyse soil samples from these fields (**Table 2.6**).

Table 2.6. Chemical characteristics of two sampled field soils

Characteristic	Field 1	Field 2
Textural class	Sandy loam	Sandy silt loam
pH	6.2	6.4
Soil density (g/L)	1118	1136
Sand 2.00-0.063mm (%w/w)	61	40
Silt 0.063-0.002mm (%w/w)	29	47
Clay <0.002mm (%w/w)	10	13
Conductivity (µS/cm)	2332	2505
Available iron (mg/kg)	39.9	69.5
Available manganese (mg/kg)	11.8	12.5
Available phosphorus (mg/kg)	48.3	60.7
Available potassium (mg/kg)	209.3	400.5
Available magnesium (mg/kg)	41.1	93.3
Available copper (EDTA) (mg/kg)	2.1	4.5
Available zinc (EDTA) (mg/kg)	2.3	6.2
Available sulphate (mg/kg)	30.4	64.4

The texture of the fields was loam to sandy loam, which corresponds to rather high levels of sand and low levels of clay with variable levels of silt (**Table 2.6**). The soils had an average pH of 6.3, which is usually associated with a high diversity of microbes (Fierer and Jackson, 2006). The concentrations in phosphorus, magnesium and iron were average to high, possibly because of the exogenous addition of fertiliser by the farmers.

2.8.2.3. Artificial contamination of leaves by E. coli

Spinach (*Spinacia oleracea cv. Picasso*) plants were grown for day/night cycles of 10 h at 20°C and 14 h at 15°C (75% constant humidity) in controlled environment rooms

by the John Innes Centre Horticultural Services staff (Norwich, UK). Four weeks-old plants were transferred to the laboratory and sprayed with a 10^6 cells/ml suspension of bacteria (*E. coli* or *Salmonella*) using an artist's airbrush (cat: HUM300133; equipped with 400ml Airbrush Pressure Bottle; cat: HUMAIR400; Humbrol, eModels) for fine and even particle deposition on the leaf surface. Plants were then incubated until sampling at 22°C in a heated propagator (Vitopod Propagator, cat: 579519; Suttons) and watered regularly from below.

2.8.3. Extraction of environmental DNA

2.8.3.1. DNA extraction from soil

DNA was extracted from soil using the FastDNA® SPIN Kit for Soil (cat: 6560-200; Q-Biogene). The major constraint to a good DNA extraction from soil is the efficient lysis of bacteria. Indeed, soil is highly heterogeneous with multiple microhabitats potentially protecting living cells, and needs to be mechanically homogenised prior to cell lysis. Additionally, many bacteria from environmental sources themselves are likely to be recalcitrant to lysis too, as they are likely to have to cope with multiple environmental stresses, or even be present in soils as stress-resistant spores. To ensure good lysis and minimal shearing of DNA molecules, the FastDNA® SPIN Kit for Soil employs a bead-beating procedure using a high-speed "up-and-down" shaker (FastPrep® FP220A Instrument; cat: 6001-220; Q-Biogene), in order to homogenise soil particles. Briefly, up to 500 mg of soil were weighed and placed into a "Lysing Matrix" tube consisting of sterile ceramic and silica beads of multiple sizes. Buffers were added (978 µl of "sodium phosphate buffer" and 122 µl of "MT buffer", both

provided in the kit) for good homogenisation and protein solubilisation. Samples were shaken for 40 s at a speed of 6 m/sec using the FastPrep® instrument before being placed on ice and centrifuged for 1 min (speed was 14,000 × *g* for every centrifugation step in this protocol). Supernatants were carefully transferred to tubes containing 250 µl of "protein precipitation solution (PPS) reagent" from the kit, and solutions were mixed by gently inverting 10 times. Tubes were then centrifuged for 5 min to pellet the protein precipitate. Supernatants were carefully transferred to 15 ml Falcon-tubes containing 1 ml of "binding matrix suspension", consisting in a mixture of fast-sedimenting mineral particles on which DNA will adsorb. Tubes were manually inverted for 2 min and sedimentation was allowed by placing tubes in a vertical position for 5 min. Some of the supernatant (500 µl) was transferred to a clean tube, and the sedimented matrix was resuspensed in the remaining supernatant (this step was meant to reduce the volume for the following steps). After the mix appeared visually homogenous, 700 µl were transferred to a membrane-based spin column and centrifuged for 1 min. The flow-through was discarded and the rest of the supernatant was passed through the membrane by centrifuging for 1 min. A washing solution (500 µl of "salt/ethanol wash solution", or "SEWS-M" from the kit) containing ethanol was added to the spin column, which was centrifuged for 1 min and, after discarding the flow-through, for an additional 2 min to dry the membranes. Tubes were dried for an additional 5 min at room temperature and membrane-bound DNA was eluted in 100 µl of nuclease-free water by centrifugation for 1 min. Samples were checked by low-voltage (60V) electrophoresis on a 0.7% TAE-agarose gel. Electrophoresis was preferred as compounds co-purified using this method can be detected by Nanodrop spectrophotometry, leading to false concentration estimates.

2.8.3.2. DNA extraction from leaves

The protocol described above was modified for leaf material homogenisation and the extraction of DNA from the phyllosphere. About 15 g of leaf material were placed in a stomacher filter bag (cat: W40545; Fisher Scientific) containing 50 ml of sterile water and crushed for 2 min at max speed in a lab blender (cat: MPR-410-012Q; Seward Medical). The filtered liquid containing bacteria was centrifuged at low speed (2 min at 500 × *g*) to pellet plant particles. The rest was centrifuged for 20 min at 13,200 × *g* and the resulting supernatant was discarded. Freezing samples at -20ºC after this point did not affect subsequent DNA purification yield. The pellet was carefully transferred in a "Lysing Matrix" tube of the FastDNA® SPIN Kit for Soil containing "sodium phosphate buffer" and "MT" buffers as described above. DNA extraction was performed as described above for soil.

2.8.3.3. PCR from environmental DNA

We used DNA extracted from environmental sources (soil and leaves) as a template for PCR targeting fragments of the 16S ribosomal RNA (rRNA) gene. This gene, encoding a component of the 30S bacterial ribosome, is quasi-universally conserved in bacteria and archaea (Woese and Fox, 1977). Its structure is not uniform, with conserved regions and more flexible, so-called hypervariable regions (Neefs et al., 1990). The sequence in these hypervariable regions is considered to be relatively species-specific. There are 9 hypervariable regions in the 16S rRNA gene, which can all be targeted by universal primers (**Figure 2.12**).

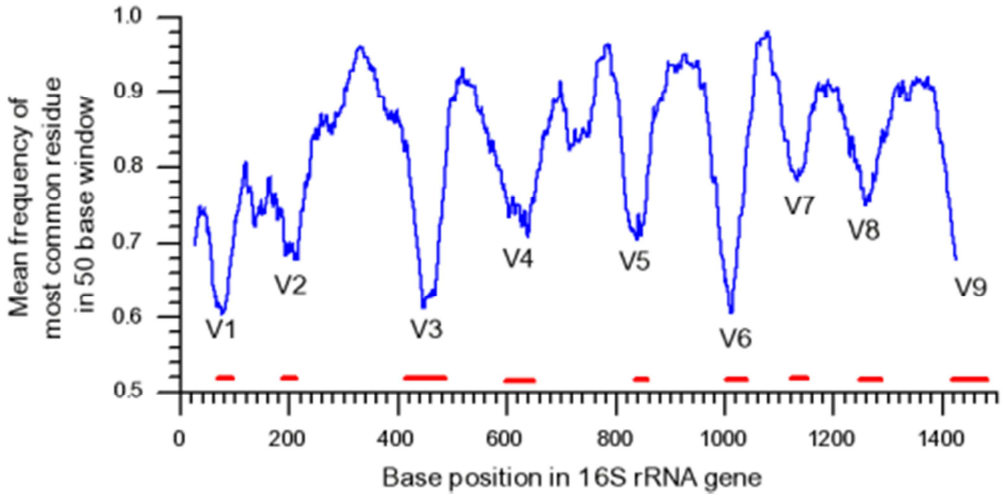


Figure 2.12. Representation of hypervariable regions within the prokaryotic 16S rRNA gene. The plot represents the conservation of bases within bacteria: high values represent conserved bases and low values represent variable bases. *E. coli* rDNA base positions are taken as reference in the x-axis. This figure was taken from the Bioinformatics Toolkit website of Cardiff University, www.bioinformatics-toolkit.org/Help/Topics/hypervariableRegions.html created by K.E. Ashelford (last accessed on 13-09-2011).

For method development purposes, we also used primer pairs 63F-338R targeting the V1-2 region, 63F-534R targeting the V1-3 region, 341F-907R targeting the V3-5 region and 1055F-1406R targeting the V7-8 region (**Table 2.7**). In our work, we focused on the utilisation of universal primers 341F and 534R (Muyzer et al., 1993) targeting the V3 hypervariable region of the 16S rRNA gene (**Table 2.7**).

Table 2.7. Universal primers used for DGGE. A 40-bp GC-clamp was added at the 5' end of primers in bold (see text for more details)

Name	Sequence (5' > 3')
63F	CAGGCCTAACACATGCAAGTC
338R	GCTGCCTCCCGTAGGAGT
534R	ATTACCGCGGCTGCTGGCA
341F	CCTACGGGAGGCAGCAG
907R[a]	CCGTCAATTCMTTTGAGTTT
1055F	ATGGCTGTCGTCAGCT
1406R[a]	ACGGGCGGTGTGTRC

[a] Standard nucleotide codes; M=A or C and R=A or G.

V3 is the most commonly used variable region for discriminating between bacterial species from various natural environments (Kadivar and Stapleton, 2003; Yu and Morrison, 2004). A 40-bp GC-clamp (5'-CGCCCGCCGCGCCCCGCGCCCGT CCCGCCGCCCCCGCCC-3') was added at the 5' end of primers written in bold in **Table 2.7** to allow immobilisation of the two strands at the same position on a denaturing gel (see section 2.8.4).

Reactions and amplification were prepared using the following parameters: for a 50-µl reaction, 25 µl of 2X GoTaq Colorless Master Mix (Promega), 0.5 µl of 10 mM of each primer, and 5 µl of gDNA extraction. The manual of the DNA extraction kit used above states that the purified DNA can be readily amplified by PCR. However, we observed difficulties in doing so, and found that the addition of 5 µg/µl of bovine serum albumin (BSA; cat: 10711454001; Roche Applied Science) in the PCR mix greatly improved the amplification success (Kreader, 1996). Amplification conditions were as follows: 1 cycle at 94°C for 5 min; 35 cycles at 94°C for 20s, 50°C or 55°C for 20s, 72°C for 20s; 1 cycle at 72°C for 5 min.

2.8.4. Denaturing gradient gel electrophoresis (DGGE)

DGGE is an electrophoresis-based method using polyacrylamide gels to discriminate DNA fragments differing by sequence content (indirectly from GC content) but not sequence length (Fischer and Lerman, 1979; Fischer and Lerman, 1980). The principle rests on the creation a gradient of denaturing conditions in an 8% polyacrylamide gel using urea and formamide. DNA samples to test are amplified by PCR with specific primers that add a long additional sequence of GC bases at the 3' end of the fragments. This so-called "GC clamp" prevents the complete denaturation of the DNA molecule. While double-stranded clamped DNA samples migrate along the electric current, denaturing conditions increase proportionally with the denaturing of the two strands. When the whole sequence but the GC clamp is denaturated, the migration stops. Fragments with different sequences with different contents in GC bases, will therefore stop at different migration fronts. Typically, DGGE can be performed on 16S rDNA amplicons from environmental DNA extractions, thus creating specific and comparable community profiles. We used the Ingeny PhorU 2x2 DGGE apparatus (GRI Molecular Biology) according to the manufacturer's recommendations and Stefan Green's comprehensive guide to DGGE (http://sites.google.com/site/stefanjgreen/sd, last accessed 13/09/2011).

Briefly, a gel cassette was vertically assembled with between glass plates carefully washed with ethanol and wiped for dust. A 32-samples comb was inserted between the two glass plates. Two reservoirs were loaded with 25 ml of polyacrylamide containing distinct concentrations of denaturing agent, 80µl of 20% (w/v) ammonium persulphate (APS) and 8 µl of tetramethylethylenediamine (TEMED) were added to

each reservoir. A peristaltic pump created a vertical denaturing gradient in the gel cassette. We observed that gradients of 59% to 40% and 59% to 44% provided a good range of denaturing conditions to discriminate most of the amplicons of our study. These were created by mixing appropriate amounts of pre-made polyacrylamide solutions containing 70% of urea/formamide or not (considered as “0%”). After the gel has polymerised (60 to 90 min), a stacking gel (7ml “0%” polyacrylamide, 80 µl of loading dye, 80µl of 20% APS, 8 µl of TEMED) was quickly added on top with a syringe. After polymerisation of the stacking gel, the comb was removed and the gel cassette transferred to the electrophoresis tank containing 15 L of 0.5X TAE buffer at 60°C. After loading samples, electrophoresis was performed at 80V (50mM, 10W) for 17.5 h (or at the appropriate voltage/time to deliver V×H=1400 with constant amperage). After migration, gels were transferred in 300 ml of 0.5X TAE containing 10 µl of SYBR® Green I (cat: S9430; Sigma-Aldrich) for DNA staining, and washed in 300 ml of distilled water for 5-10 min. Gels were scanned in a Pharos FX Plus Molecular Imager (Bio-Rad).

2.8.5. Data analysis

We used Phoretix 1D (Nonlinear Dynamics) demo version to analyse denaturing gradient gels. A very good description of the principle behind gel analysis using Phoretix 1D or similar softwares has been published recently (Tourlomousis et al., 2010). We summarised this process in **Figure 2.13**.

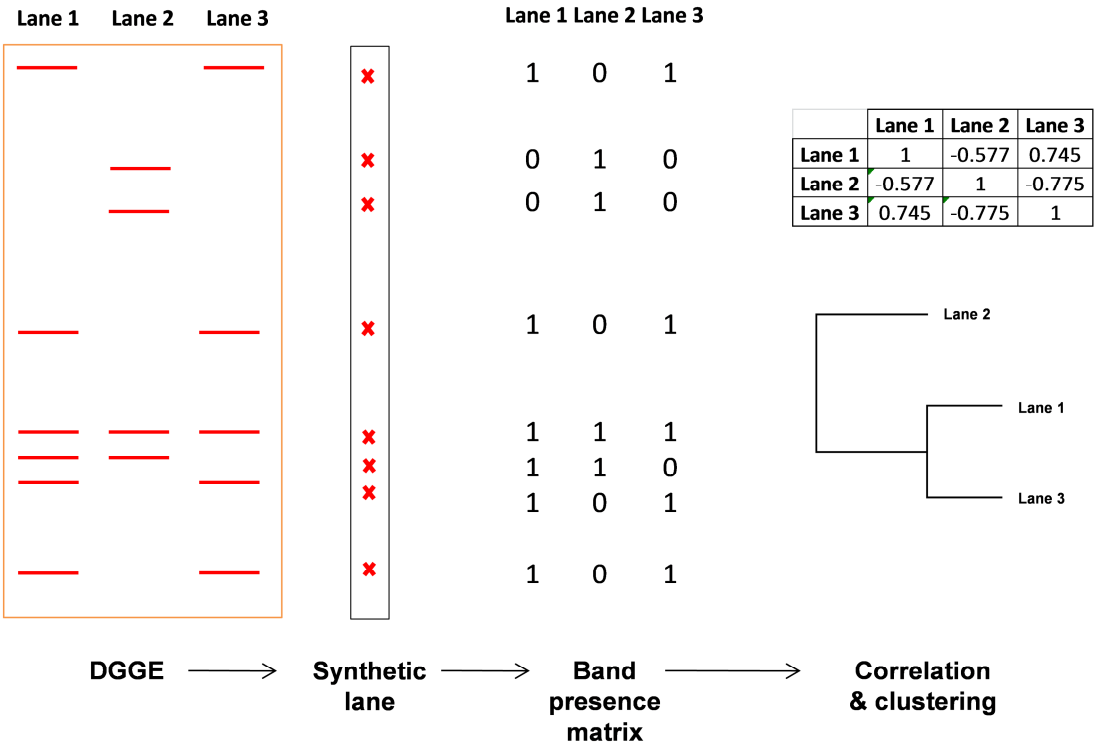


Figure 2.13. Summary of the data analysis process using gel analysis softwares. See text for more details.

Briefly, gel pictures are loaded and lanes are created. The gel background is substracted and bands are detected. This step needs to be visually controlled and edited as sometimes, the software either misses or adds unwanted bands (specks, poor gel images, etc). Then, the user defines bendable migration front lines (or “Rf lines”). This step is crucial to the analysis as it is used as a frame by the software to match similar bands and consider different bands as such. The output of this analysis is first a synthetic lane regrouping all different band positions of one gel, from which a band presence matrix is created. From this, Pearson correlations can be calculated on the profiles, and clustering of correlated profiles can be performed (**Figure 2.13**).

3. Genomic and phylogenetic diversity of *E. coli* strains seasonally isolated from agricultural crops

3.1. Context

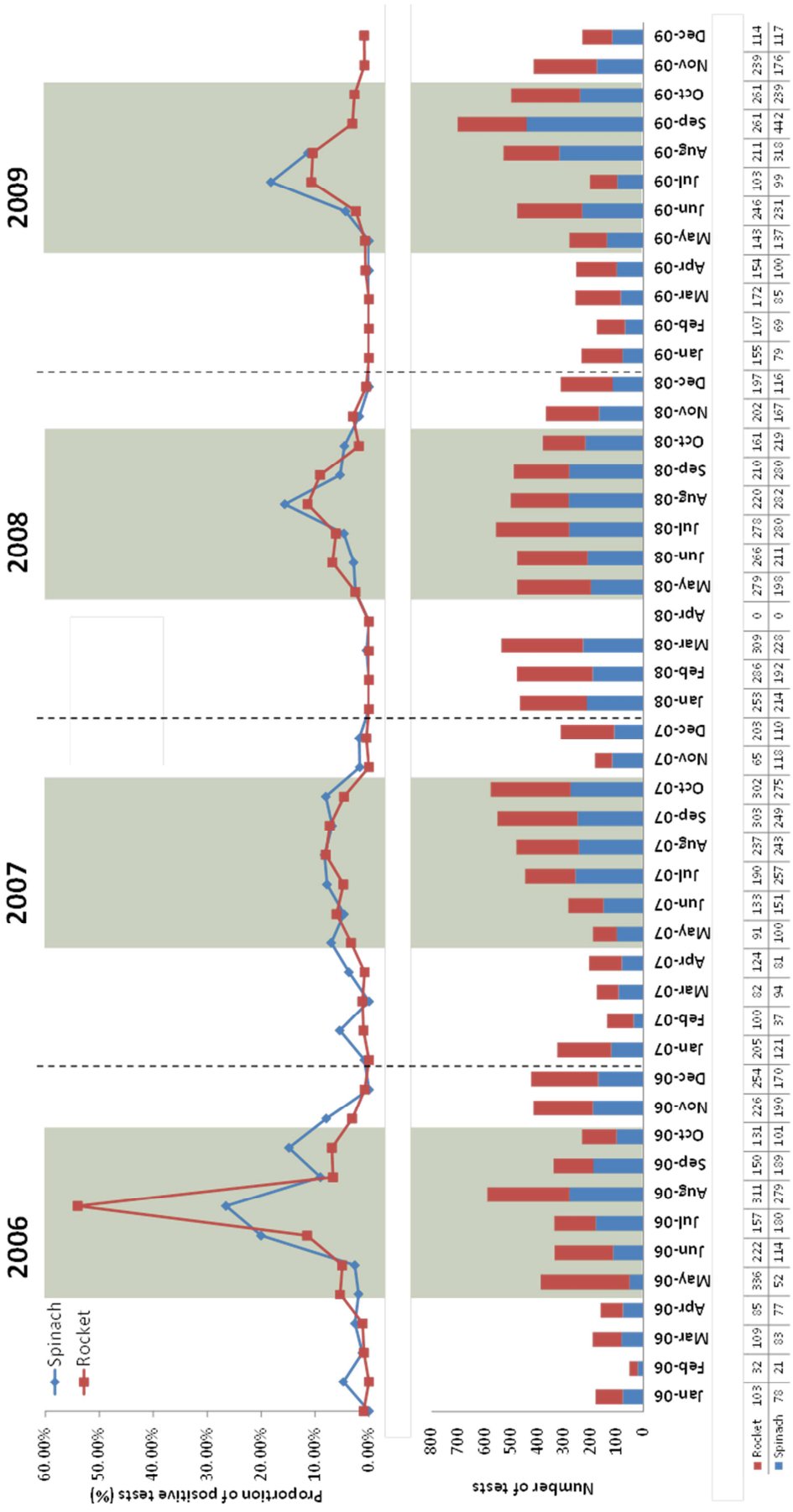
Since the first analyses of the *E. coli* population structure and the resulting observation of distinct phylogenetic groups, hypotheses have emerged as to why these phylogroups exist, and why different distributions are observed in collections of isolates of different origin. The question of how the living environment acts on the distribution of phylogenetic groups, and thus the population structure of a particular bacterial species remains poorly understood. These questions can be extended to our topic of study. Are nonhost secondary environments acting on the population structure of *E. coli* GMB isolates? Are strains of *E. coli* isolated in nonhost secondary environments a specific subset of all *E. coli* or do they encompass the whole diversity of the species? In other words, is nonhost adaptation a part of the natural life cycle of *E. coli*? Agricultural fields are arguably very good nonhost environment models. *E. coli* is not believed to be commonly associated with plants, although it has been shown to survive and persist for relatively long times in agricultural environments (see Introduction section 1.3.1). Therefore, when *E. coli* strains are isolated from the aerial parts of plants without any obvious source of contamination, it is possible that they do not come directly from faecal matter, but already survived the selection pressures, if any, of other nonhost environments before, such as water or soil. In this section, we first present data showing that *E. coli* can be seasonally isolated in abundance during planting seasons from the leaves of salad crops. Based on this information, we hypothesised a meteorological effect on the seasonality of *E. coli* spikes of detection on plants. Then, to characterize further the association of

environmental *E. coli* with plants, we compared isolates from distinct sources, symbolised by two collections of strains (GMB and ECOR). The ECOR collection is composed of strains associated with human and animal hosts (primary environment), whereas GMB strains isolated in this study come from the agricultural environment, more precisely from salad crops and soil (secondary environment). We first examined the distribution of the phylogenetic groups of GMB isolates and decomposed it according to various known parameters about these strains to see how they could influence the abundance of specific phylogroups. We then compared the diversity between ECOR and GMB strains, first at the genomic structure level (using a PCR fingerprinting method derived from REP-PCR) and then at the phylogenetic level (using MLST).

3.2. Seasonality of *E. coli* contamination of crops

3.2.1. *E. coli* environmental isolates are commonly isolated from salads growing in the UK

The observation that led to the initiation of this project was the finding that, after routine microbiological safety tests conducted by anonymous industrial partners, *E. coli* (monitored as a faecal indicator and called "generic *E. coli*" in food safety regulations) could be isolated from UK-grown salads during the growing season (approx. May to October). It was also observed that the quantity of *E. coli* retrieved from all the tests on spinach and rocket was varying greatly from one year to another (**Figure 3.1**). It is worthy to note that the data presented are a compilation of various growers for the same crop, and that no single geographical location was identified as an increased source of *E. coli*.


2006
2007
2008
2009
Spinach
Rocket
Proportion of positive tests (%)
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Number of tests
800
700
600
500
400
300
200
100
0
Jan-06 Feb-06 Mar-06 Apr-06 May-06 Jun-06 Jul-06 Aug-06 Sep-06 Oct-06 Nov-06 Dec-06 Jan-07 Feb-07 Mar-07 Apr-07 May-07 Jun-07 Jul-07 Aug-07 Sep-07 Oct-07 Nov-07 Dec-07 Jan-08 Feb-08 Mar-08 Apr-08 May-08 Jun-08 Jul-08 Aug-08 Sep-08 Oct-08 Nov-08 Dec-08 Jan-09 Feb-09 Mar-09 Apr-09 May-09 Jun-09 Jul-09 Aug-09 Sep-09 Oct-09 Nov-09 Dec-09
Rocket 103 32 109 85 336 222 157 311 150 131 226 254 205 100 82 124 91 133 190 237 303 302 65 203 253 286 309 0 279 266 278 220 210 161 202 197 155 107 172 154 143 246 103 211 261 261 239 114
Spinach 78 21 83 77 52 114 180 279 189 101 190 170 121 37 94 81 100 151 257 243 249 275 118 110 214 192 228 0 198 211 280 282 280 219 167 116 79 69 85 100 137 231 99 318 442 289 176 117

←**Figure 3.1. Detection of *E. coli* on spinach and rocket field-grown in UK and overseas during routine microbiological safety tests**. The corresponding number of total tests for rocket and spinach is shown in the columns and table below the line plot. Areas shaded in dark correspond to the growing seasons in the UK (May to October of each year) Areas in white correspond to produce grown overseas (typically Spain or Italy), shipped to UK for microbiological testing and commercialisation. This figure was created using data from anonymous industrial partners.

Between 2006 and 2009, routine tests lead to the detection of *E. coli* in 958/16,827 (5.69%) tested spinach and rocket samples. There was a massive detection rate of *E. coli* on spinach and rocket during summer 2006, with a peak in August 2006 where 74/279 (26.52%) samples of rocket and 168/311 (50.16%) samples of spinach tested positive for at least 10 CFU of *E. coli* per gram of tested leaves. In the 4-year period investigated, the overall distribution of positive tests for spinach and rocket was strongly correlated (Spearman correlation; r_S=0.8187, p<0.0001), even if only considering the UK-based production (May to October) for each year (r_S=0.7223, p<0.0001).

After 2006, detection rates remained below the one observed in summer 2006, but there were detection peaks of medium importance during summer 2008 and 2009, with more than 10% of samples being positive for *E. coli* in August 2008, and in July and August 2009 (**Figure 3.1**). Also, no specific increased detection was observed during summer 2007 (**Figure 3.1**). Overall, tests performed in the UK for produce grown overseas (the white areas on **Figure 3.1**) never detected *E. coli* at comparable levels as in tests of plants grown in UK during summer.

3.2.2. *E. coli* contamination of conventional and organic salad

In the light of produce-related outbreaks by pathogenic *E. coli*, there is an industrial and societal interest on whether the type of agriculture can influence contamination by *E. coli*. Indeed, one may think that organic foods could potentially harbour more faecal organisms because of the increased use of manure for fertilisation to compensate for not using chemical fertilisers. We also collected data on *E. coli* isolation rates from rocket salad grown conventionally and organically throughout the UK (**Figure 3.2**).

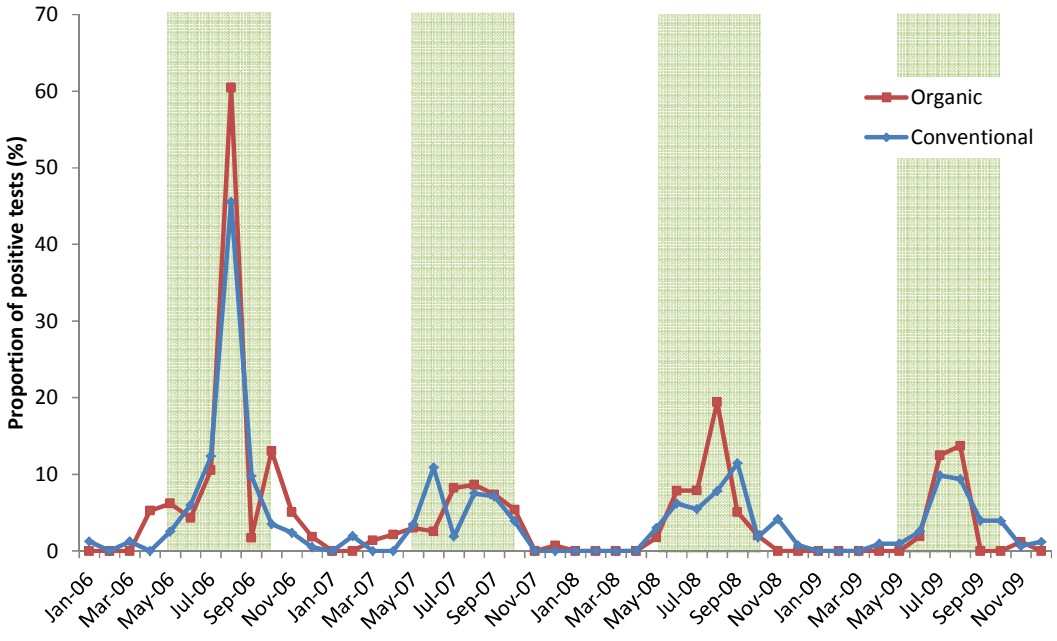


Figure 3.2. Detection of *E. coli* on conventionally and organically grown rocket from UK (green shades) and overseas (white shades) fields during routine microbiological safety tests. See legend of **Figure 3.1** for more details.

There were slightly more positive tests for *E. coli* on organically grown than conventionally grown rocket leaves during the period tested but the two distributions

were not statistically different (r_S=0.6672, p<0.0001). Similarly, when we compared the distributions of summer months only (UK-based production, May to October for each year), the correlation was still not significant (r_S=0.4578, p=0.245). From these data, we can conclude, there were no significant differences between the numbers of positive tests for *E. coli* on organic compared to conventionally grown rocket salad.

In a similar study, fresh lettuce grown organically or conventionally in Spain were examined and authors found that the levels of *Enterobactericeae* present on leaves were the biggest source of variation between organically and conventionally grown plants (Oliveira et al., 2010). In this study, more *Enterobacteriaceae* were retrieved from organic than conventional plants (Oliveira et al., 2010). Similarly a pre-harvest local farm-based study in USA showed that organic produce had a higher prevalence of *E. coli* than conventional salads, especially in farms where manure or compost aged less than 12 months had been used as a fertiliser (Mukherjee et al., 2004). On the other hand, an earlier study on UK-grown salads concluded without comparison with conventionally-grown crops that organic ready-to-eat vegetables were of very satisfactory microbiological quality (Sagoo et al., 2001). Our short analysis based on routine microbiological safety monitoring data tends to confirm that there is no difference in *E. coli* detection (and thus the inferred "microbiological safety") between conventional and organic UK-grown salads.

3.2.3. Meteorological conditions in UK from 2006 to 2009 and hypotheses for seasonality

To our knowledge, no link has ever formally been established between meteorological factors such as temperature and rainfall and *E. coli* contamination of agricultural

crops in fields. Additionally, our systematic *E. coli* detection dataset covers multiple locations in the UK and thus allows for a more general analysis regarding country-wide variations in *E. coli* detection rates on spinach and rocket, rather than sporadic contamination events. It was observed that the total microbial load on minimally processed spinach was higher in summer and autumn than on spring and winter, suggesting that temperature and rainfall variation impact greatly on bacterial growth on plants (Caponigro et al., 2010). Additionally, fast-growing bacteria such as *E. coli* tend to grow to higher levels at higher temperatures in laboratory conditions.

We then looked at the correlation between recorded meteorological conditions from 2006 to 2009 (obtained on the website of the UK Meteorological Office, or "MetOffice", http://www.metoffice.gov.uk/climate/uk/; last accessed: 28/09/2011) with *E. coli* detection rates on crops based on data shown in the previous parts. We first looked at average anomaly data for the whole of UK. Here, the anomaly is defined as the observed difference with the average values for all years between 1971 and 2000.

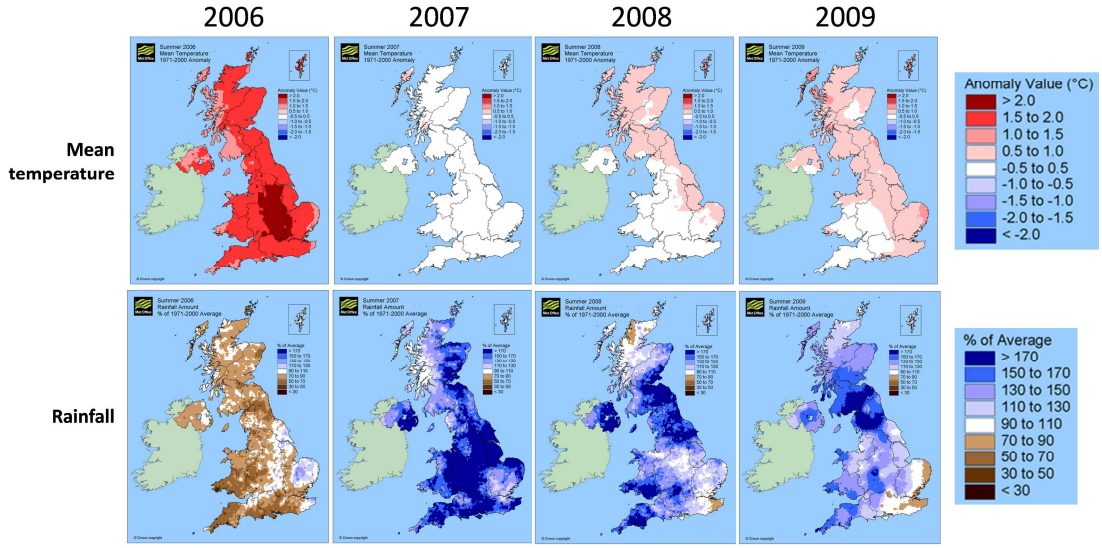


Figure 3.3. Average in anomaly (1971-2000) of mean temperature and rainfall for Summer (June to September) 2006 to 2009 in the UK. Colours are defined in the legends on the right, and correspond to the difference observed from the average of all values between 1971 and 2000 for the year mentioned on top. Data presented here is subjected to a crown copyright by MetOffice.

Overall, the variations in anomalies for mean temperature and rainfall for summers 2006 to 2009 followed three different profiles (**Figure 3.3**). In 2006, temperature was high in summer, and rainfall was very low (with the exception of Norfolk). In 2007, temperature was average, but rainfall was very high. Similar slightly higher temperature and higher rainfall were observed in 2008 and 2009. Using distributions for all months between 2006 and 2009, we found that for the whole of the UK the pattern of *E. coli* detection on spinach or rocket salad was strongly correlated with the temperature evolution from 2006 to 2009 (r_S=0.8275, p<0.0001), unlike rainfall (r_S=0.1743, p=0.2361). When we used only the months of UK-based production (May to October), similar correlations trends were obtained for temperature (r_S=0.6368, p=0.0008), and rainfall (r_S=0.1435, p=0.5036).

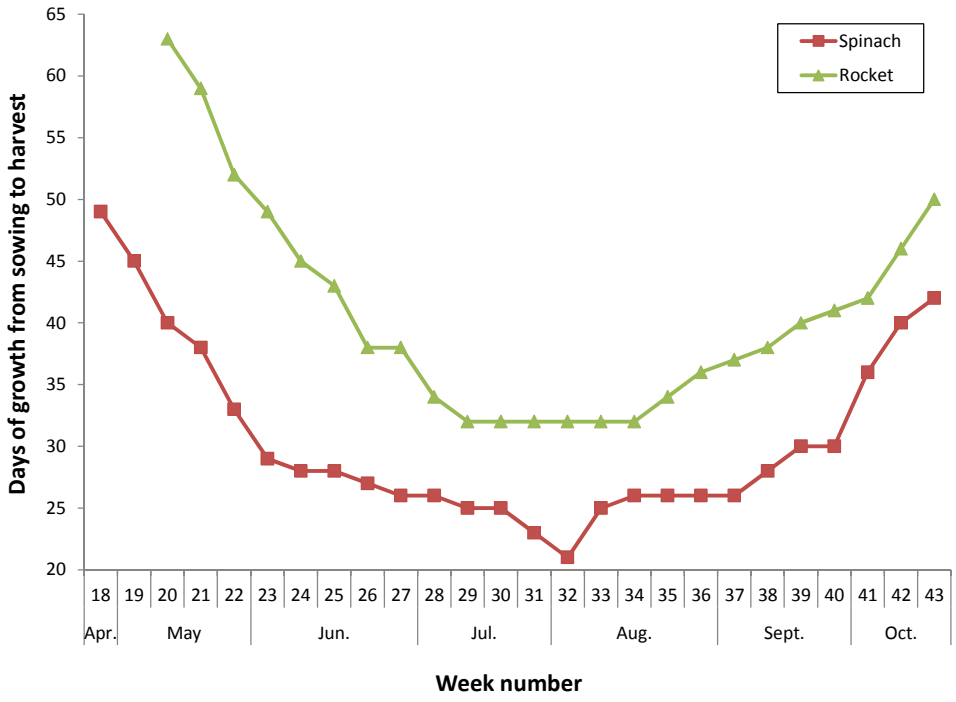


Figure 3.4. Number of days from sowing to harvest according to the time of the year for spinach (red) and rocket salad (green). Data courtesy of anonymous farm partners.

There is a correlation between temperature and the retrieval of *E. coli* from agricultural plants. *E. coli* primary habitat being the warm-blooded animal gastrointestinal tract, it is better adapted to grow at higher temperatures than the environmental average, and it is plausible that a higher temperature simply acts on the growth rate of bacteria living on plants, with more *E. coli* being retrieved during the warmer months of the year. The correlation could also be indirect, as spinach and rocket cultivars used commercially can grow around 30 days faster in July/August than in April/May (**Figure 3.4**). The summer photoperiod being longer during summer days, photosynthesis produces more energy, which is transported within the plant by sucrose, possibly affecting bacterial growth. Similarly, a more rapid growth of vegetables could produce more nutrients leakage, and an overall higher concentration of nutrients available for *E. coli* to colonize.

3.3. Description of a collection of *E. coli* isolates from plants ("GMB" collection) and their phylogroup distribution

Our first insight into *E. coli* association with plants of agricultural interest indicates that *E. coli* association with plants is not uncommon, as reported previously by the literature (Mandrell, 2009). In our opinion and from an ecological point of view, agricultural plants constitute a very likely secondary habitat (or nonhost environment) for *E. coli*. Sources of salad contamination by enteric bacteria are likely to be multiple and complex. Direct faecal contamination of plants by wildlife faecal deposition is possible, but the most likely ways of field contamination are through the spreading of manure in soils, and irrigation (Brandl, 2006). One can then imagine that the history of various environmental strains isolated from plants is likely to be very diverse and shaped by the multitude of different nonhost environments these strains have persisted in before their arrival in the phytosphere. The examination of intra-species diversity in nonhost environmental isolates of *E. coli* retrieved from plants is the precise focus of most of the work presented in this PhD thesis, with this present section being on the basic phylogenetic characterisation of the strains.

3.3.1. Combination of triplex PCR and MLST to determine phylogroups

Since the first observation of the clonal structure of *E. coli* natural population, various methods have been employed to assign phylogenetic groups to environmental isolates. Multiple loci-based methods such as MLVA or MLST have been proved to be the most accurate and powerful in discrimination. However, they remain time-consuming and laborious. The utilisation of 3 phylogenetic markers detectable by triplex PCR,

whose patterns of presence can accurately predict for one of the 4 major phylogroups, A, B1, B2 or D, was proposed (Clermont et al., 2000). This accuracy was verified by comparing assignments using the triplex method and MLST on a large number of isolates (Gordon et al., 2008). In this work, we used the triplex method to assign all GMB isolates (n=106) to the 4 major phylogroups, and we constructed phylogenetic relationships for a subset of them (n=76, 71.7% of the full collection) based on MLST data. The attribution of phylogroups with MLST was inferred from the clades in the phylogenetic tree in **Figure 3.18**. The distribution of phylogroups among GMB isolates as observed using the triplex PCR method can be seen on **Figure 3.5** (column 1).

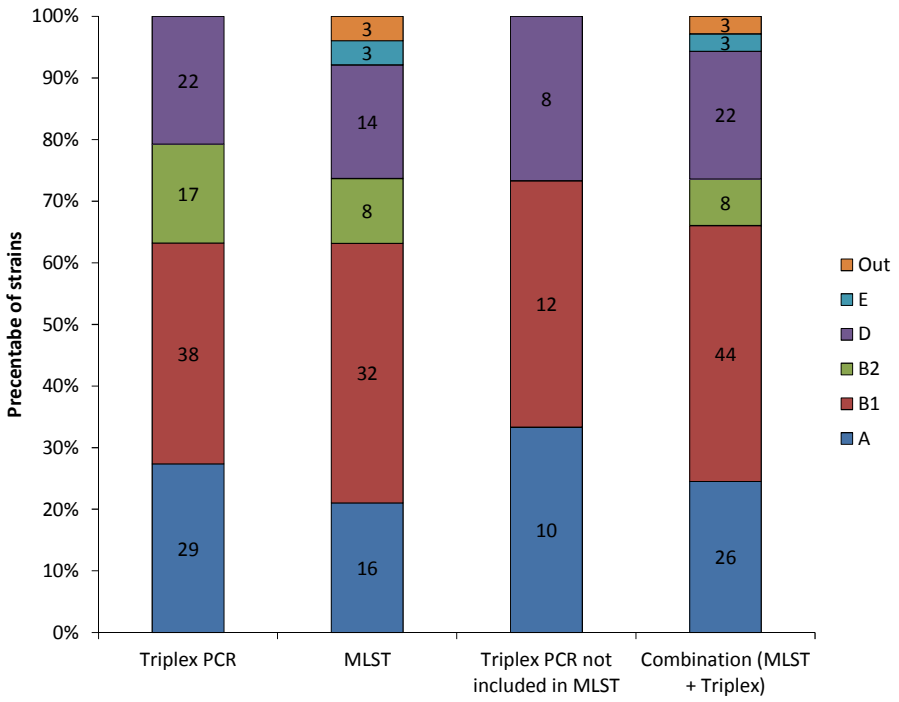


Figure 3.5. Phylogenetic groups in *E. coli* strains isolated from plants (GMB). Phylogroup distribution according to: triplex PCR method on all GMB isolates (column 1), MLST on a subset of GMB isolates (column 2), triplex PCR on isolates that were not included in the MLST analysis (column 3) and a combination of column 2 and 3 (column 4). "Out" encompasses isolates from cryptic *Escherichia sp.* clades (see section 3.4.2.2).

As we limited the number of isolates tested by MLST for cost reasons, there are isolates that we did not include in the MLST analysis (**Figure 3.5**, column 3), which are mostly from phylogroups A, B1 and D to minimize the bias of not taking into account those strains (**Figure 3.5**, column 2). Additionally, the determination of phylogroups using triplex PCR (**Figure 3.5**, column 1) and MLST (**Figure 3.5**, column 2) were not statistically different (χ^2=1.755; p=0.6247), stressing once more the reported accuracy of the triplex PCR method. We therefore combined (**Figure 3.5**, column 4) the results of phylogroup assignment according to MLST and to the triplex PCR method, for isolates not included in the MLST experiment. All distributions presented after **Figure 3.5** are based on this distribution. Among GMB strains, phylogroup B1 was the most prevalent (44/106, 41.5%). Phylogroups A and D were similarly represented (26/106, 24.5% and 21/106, 19.8% respectively) whereas strains from phylogroup B2 were in the clear minority (9/106, 8.5%**)**. Interestingly, 3 isolates clustered with phylogroup E, which is only identifiable by MLST (triplex PCR showed the corresponding isolates to belong to phylogroup B2). Additionally, 3 isolates (GMB46, 56 and 57) were also not clustering within the habitual phylogenetic groups of *E. coli* (although triplex PCR assigned them to phylogroup B2) and were labelled "Out" for "outgroup". A more thorough analysis of the outgroup is presented in section 3.4.

3.3.2. Phylogroup distribution according to various parameters

GMB strains were isolated from various geographical locations, on different salad crops and times during the year. In the following sections, we examined the

distribution of phylogenetic groups according to these different parameters to provide the reader with an accurate representation of the major variables within the 106 strains of the GMB collection.

3.3.2.1. Plant of isolation

GMB isolates were isolated mainly from spinach (*Spinaciaoleracea*) and rocket (*Erucasativa*) plants, but also from other salad crops, mixtures of leaves (in salad bags) or field soil. **Figure 3.6** shows the distribution of phylogroups in strains from these various origins of isolation.

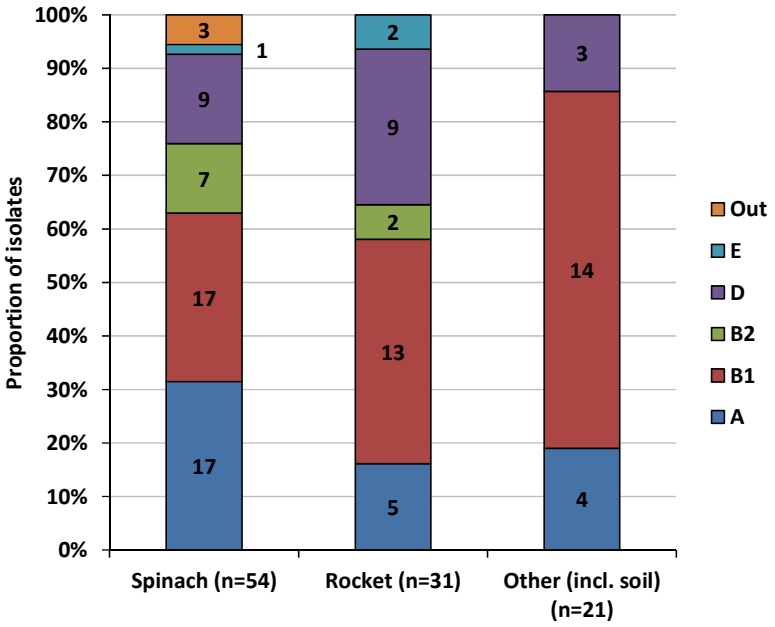


Figure 3.6. Proportions of *E. coli* phylogenetic groups in the GMB collection of isolates according to plants of isolation. Numbers in columns represent the number of isolates for each phylogroup.

The 4 major phylogroups A, B1, B2 and D were all present in isolates from spinach (**Figure 3.6**, column 1). The phylogroup distribution among isolates from spinach was strongly correlated to the phylogroup distribution of all GMB strains (r_S=0.9706, p=0.0028). All 3 “outgroup” strains were found on spinach, and strains from

phylogroup E were found on both spinach (1 strain) and rocket (2 strains). Notably, a smaller proportion of strains from phylogroup A and B2 were found in isolates from rocket, and a larger proportion of D (**Figure 3.6**, column 2), but the phylogroups of isolates from rocket were still representative of all GMB strains (r_S=0.8971, p=0.033), as were isolates from sources other than spinach and rocket (r_S=0.9549, p=0.0167). Phylogroup distributions in isolates from spinach and rocket were nevertheless not statistically similar (r_S=0.7941, p=0.058) but those in isolates from spinach and other types of salads were (r_S=0.9241, p=0.0167). These correlations could just be caused by the strong bias in sampling towards spinach and rocket-associated isolates (i.e., what is observed in spinach and rocket is similar to what is observed in the whole), but may also indicate that the plant of origin has little effect on the phylogroup distribution of colonizing *E. coli* strains.

3.3.2.2. Geographical location

To investigate the possibility of a geographical effect on the population structure of GMB isolates, we decomposed the distribution of phylogenetic groups according to different parameters. Among 30 isolates from rocket, 22 (73.3%) were isolated on the same day in the same field (GMB59 to 81), potentially explaining why we previously did not observe a statistical correlation between phylogroup distributions in isolates from spinach and rocket (**Figure 3.7**).

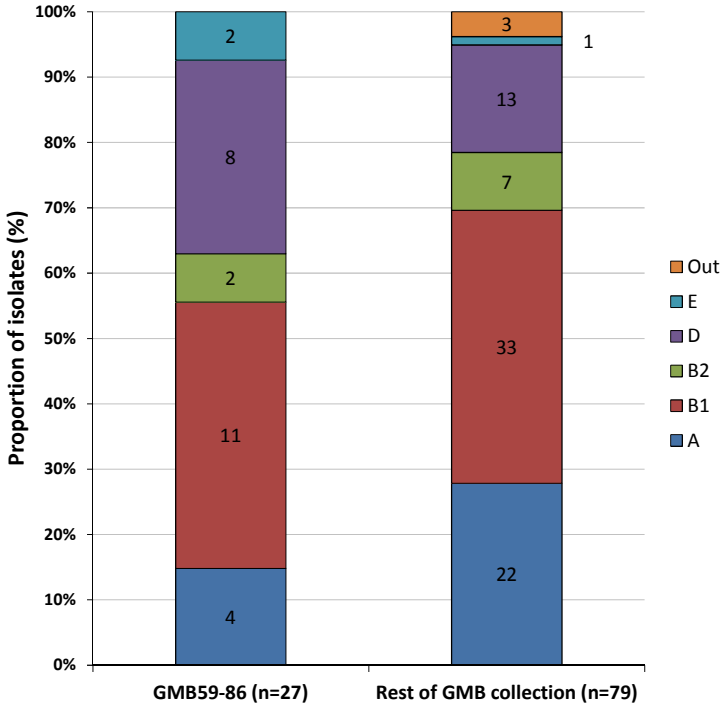


Figure 3.7. Proportions of *E. coli* phylogenetic groups in strains isolated from one rocket field at the same time, compared to the rest. Numbers in columns represent the number of isolates for each phylogroup.

The phylogroup distribution of these field-specific isolates from rocket salad was obviously correlated to the distribution of all rocket isolates (r_S=1, p=0.0028), but also to the whole GMB phylogroup distribution regardless of the plant of isolation (r_S=0.8971, p=0.0333), suggesting that there is no "field effect". In other words, GMB59 to 81, isolated on the same day in a rocket field, were overall good representatives of the phylogroup diversity we could find in our whole GMB collection of isolates from plants.

3.3.2.3. Time of isolation

GMB strains were isolated in 2008 and 2009 during the salad planting season in the UK (usually from May/June to September/October depending on meteorological conditions). When sampling size was large enough, rather similar phylogroup distributions were observed for the various times of isolation (**Figure 3.9**).

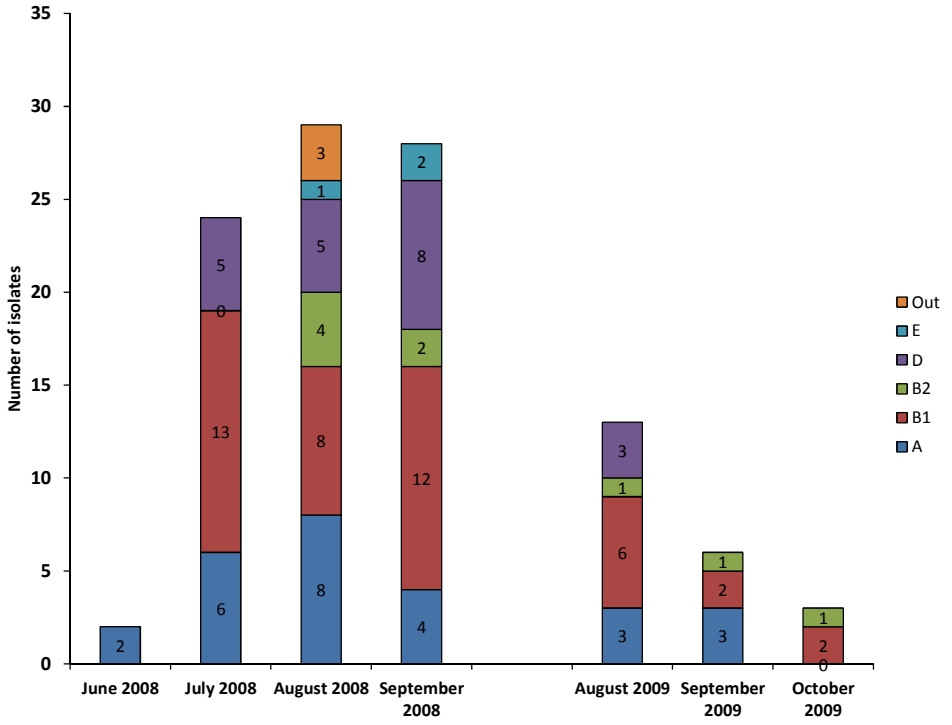


Figure 3.9. Phylogenetic groups in *E. coli* strains isolated at different times of 2008 and 2009. Numbers in columns represent the number of isolates for each phylogroup.

Isolates from June 2008, September and October 2009 were not abundant enough for any statistical significance. However, phylogroup distributions of strains isolated in July (r_S=0.9549, p=0.0167), August (r_S=0.9706, p=0.0028), September 2008 (r_S=0.8971, p=0.0333) and August 2009 (r_S=0.9852, p=0.0028) were all statistically correlated to the distribution of all GMB isolates. Additionally, a majority of GMB isolates were isolated during summer 2008 (82/105, 78.1%), and the rest during summer 2009 (22/105, 21.0%) (**Figure 3.9**). Both phylogroup distributions of 2008 (r_S=1, p=0.0028) and 2009 (r_S=0.9852, p=0.0028) isolates were statistically correlated to the phylogroup distribution of all GMB isolates. These observations strongly suggest that, within the boundaries of our study and sampling, there was no effect of isolation time on the phylogenetic structure of plant-associated *E. coli* strains. This

also suggests that, in terms of phylogroup distribution, the GMB collection is a good representation of *E. coli* strains associated with plants from one year to another.

In Chapter 4, we investigate this hypothesis by identifying traits specifically associated with plant-associated strains and specifically phylogroup B1. Working under the assumption that (a) nonhost strains isolated from plants are likely to have been in contact with nonhost environments (outside mammalian intestines), (b) traits conferring fitness in a given environment will be enriched in strains isolated from this environment and (c) traits associated with phylogroup B1 could provide an insight in to nonhost adaptation in *E. coli*, we adopted in the next parts a collection-wide comparative analyses approach, by comparing the GMB collection of plant isolates with the 72 faecal isolates of the *E. coli* Reference (ECOR) collection, meant to encompass the whole genetic variability of *E. coli* as a species (Ochman and Selander, 1984).

3.4. Diversity and phylogeny of plant-associated *E. coli* and comparison with host-associated ECOR strains

3.4.1. Clonal diversity of plant-associated *E. coli*

Our investigation to identify differences between host (ECOR) and nonhost (GMB) *E. coli* strains starts at the phylogenetic diversity level. The triplex PCR assay used in the previous sections provided insights into the population structure of our strains by indicating the distribution of their phylogenetic groups. However, there could be subgroups within the phylogroups, or even relatedness or distance between them, all of which the triplex PCR assay would not indicate. Importantly, knowing the phylogroup distributions of our strains does not indicate anything on the genetic diversity or the relationships between them. Indeed, the triplex method does not indicate whether plant isolates are composed of stable clones or if their diversity is high.

In this part, we examined the phylogenetic relationships of GMB strains and compared them to the *E. coli* reference collection (ECOR) strains using more powerful typing methods. We first compared genomic structure and large-scale recombination events using BOX-PCR, a DNA fingerprinting method. Using BOX-PCR, we could also approximately examine the strain diversity within our collections. We then used MLST, a sequence-based approach, to investigate further the relationships between strains. From MLST data, basic population genetics calculations were made to gather more information and estimates about the structure and diversity of the natural plant-associated *E. coli* population. We then constructed a

phylogenetic tree to visually assess the phylogenetic position of GMB strains within the *E. coli* species, as symbolised by the ECOR collection.

3.4.1.1. BOX-PCR to examine diversity

Using BOX A1R primers to amplify DNA sequences located between repeated regions in the genome indicates how different the genomic architecture between isolates is. BOX-PCR profiles are thought to be stable over many bacterial generations but can vary over longer time. Between two given repeats in the genome, events such as prophage insertion, plasmid insertion or recombination between repeated sequences can occur and modify the sequence size between these two repeats, which will be reflected by different migration fronts on electrophoresis gels. It is therefore possible to use BOX-PCR to observe those changes between *E. coli* strains, under the assumption that similar strains or even clones will have similar BOX-PCR profiles.

Because of its capacity to show major recombination and insertion events in bacterial genomes, BOX-PCR is also an appropriate method to compare the genomic structure of various strains in order to get a crude idea of the genomic diversity. Using the TotalLab software, we transformed electrophoretic profiles for each strain into numerical tables based on the number of bands and their migration position. The profiles of each strain were then compared by PLS-DA (**Figure 3.10**).

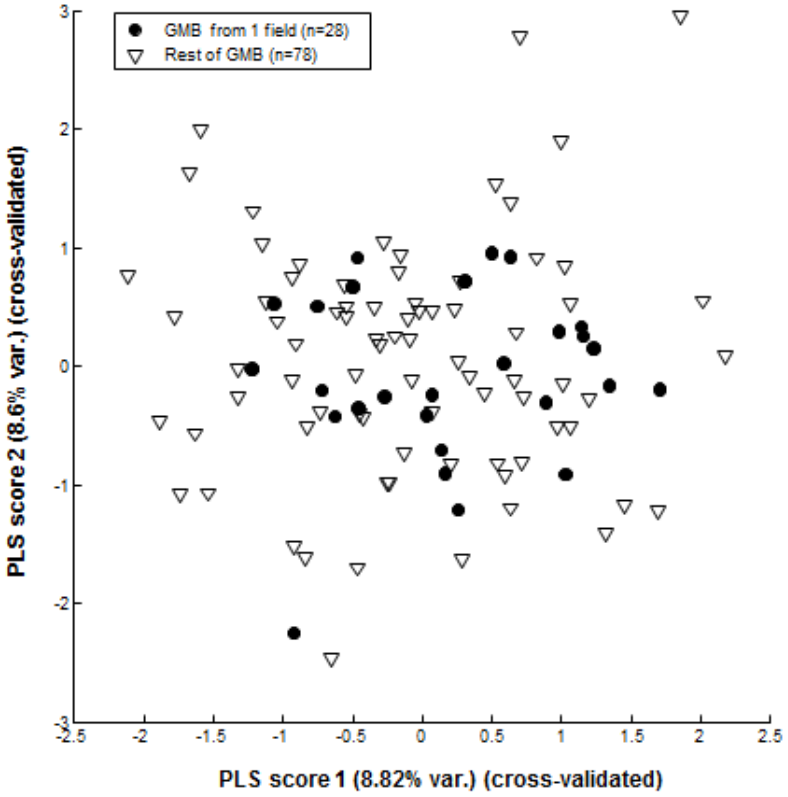


Figure 3.10. PLS-DA analysis on BOX-PCR DNA fingerprints using GMB isolates with different isolation histories. GMB isolates from the same field (n=28) are represented by filled circles; GMB strains isolated in various locations (n=76) are represented by open triangles. Success rate in PLS-DA was maximised at 2 PLS dimensions, with an insignificant cross-validation rate of 66%.

We first observed that most of the strains had different profiles, and were generally not originating from clones. GMB isolates were very diverse, and only a few of them shared identical BOX-PCR profiles (**Figure 3.10**). This could be an indication that *E. coli* contaminates plants from multiple or complex sources and that there is possibly limited growth on plants, as isolates from the same field do not generally produce similar BOX-PCR profiles neither. Indeed, isolates from September 2008, isolated on the same day, from the same field in Norfolk showed comparable genomic structure diversity as isolates from the whole of the UK (**Figure 3.10**). In other words, diversity

in one field is more or less a good surrogate for global diversity of plant-associated isolates.

Moreover, when we added ECOR strains to the analysis, we could not associate any variation specifically with either ECOR or GMB isolates (data not shown), suggesting that the genomic structure variability is similar within *E. coli* and is not associated with the source of isolation. This is expected if we consider that all *E. coli* are primarily faecal isolates and that their association with plants is not long enough to cause distinct deep environment-specific genomic rearrangements observable by BOX-PCR. As obvious as it seems, this examination was not trivial, as it was reported that "naturalised" populations of *E. coli* could be isolated from soils (see Introduction section 1.3.1) and discriminated by REP-PCR method derivatives. Our result would seem to indicate that there is no genetic isolation from plant isolates, indicating that faecal contamination, rather than long-term persistence leading to genetic isolation is occurring on plants. However, BOX-PCR is a crude way of observing genetic similarities and comparing the relatedness of isolates, which prompted us to use more powerful phylogenetic analysis methods.

3.4.1.2. Properties of the MLST scheme used in this study

Phylogenetic analysis based on the sequence of typically 6 to 8 neutrally-varying housekeeping gene sequences (regrouped in "MLST schemes") is a common way to make assumptions on the sampled population properties and phylogenetic history, providing that enough samples are sequenced and more importantly that the tested genes are indeed showing a reliable phylogenetic signal (Spratt, 2004).

To investigate any possible differences in selection patterns between collections or phylogroups, we calculated the d*N*/d*S* ratio, commonly used as an indicator of selective pressure on a particular gene (Yang and Bielawski, 2000). This calculation is based on the comparison of synonymous (d*S*) and non-synonymous (d*N*) substitution rates in specific populations of sequences: if d*N*>d*S*, the rate of non-synonymous substitutions is higher than the rate of synonymous substitutions (i.e., the protein sequence is likely to vary greatly across isolates), the ratio is superior to 1 and indicates diversifying positive selection at this locus; if d*N*<d*S*, the rate of non-synonymous substitutions is lower than the rate of synonymous substitutions (i.e., the protein sequence is likely to be very conserved across isolates), the ratio is inferior to 1 and indicates purifying (or negative) selection at this locus; if d*N*=d*S*, the ratio is equal to 1 and indicates neutral selection at this locus. An ideal MLST scheme encompasses loci that are under purifying selection, to make sure no positive selection occurs (Perez-Losada et al., 2007). Loci under neutral selection can be considered only cautiously, as an apparent neutral selection can likely be caused by an equal balance of positive and purifying selection and such a scenario would hinder the phylogenetic signal. Here, we calculated the d*N*/d*S* ratio for each locus based on 142 ECOR and GMB sequences (**Table 3.1**) in order to check if our tested strains gave expected results using the Pasteur MLST scheme.

Table 3.1.**Sequence variation at 8 loci based on 142 ECOR and GMB sequences**. Details of the calculation performed with the START2 package.

Locus	Alleles	(*S*+*N*)[a]	PC[b]	*S*[c]	*N*[d]	d*N*[e]	σ d*N*[f]	d*S*[g]	σ d*S*[h]	d*N*/d*S*
dinB	33	450	528	114.2	335.8	0.00	0.00	0.14	0.08	**0.026**
icdA	62	516	1891	122.5	393.5	0.00	0.00	0.12	0.05	**0.022**
pabB	38	468	703	112.3	355.7	0.01	0.01	0.08	0.04	**0.116**
polB	42	450	861	107.5	342.5	0.01	0.00	0.15	0.06	**0.041**
putP	49	456	1176	118.3	337.7	0.00	0.00	0.11	0.06	**0.043**
trpA	40	561	780	147.7	413.3	0.01	0.00	0.18	0.10	**0.029**
trpB	35	594	595	145.7	448.3	0.00	0.00	0.15	0.08	**0.030**
uidA	40	600	780	140	460	0.00	0.00	0.09	0.04	**0.048**

[a]Number of coding sites analysed (*S*+*N*); [b]number of pairwise comparisons made (PC); [c]mean number of synonymous sites (*S*); [d]mean number of non-synonymous sites (*N*); [e]mean non-synonymous substitutions per non-synonymous site (d*N*); [f]standard deviation of d*N*; [g]mean synonymous substitutions per synonymous site (d*S*); [h]standard deviation of d*S*.

For each locus, the d*N*/d*S* ratio was far lower than 1, indicating a strong purifying selection. The ratio for *pabB*, encoding for a *p*-aminobenzoate synthase was slightly higher than other genes. To examine if selection pressures are of similar nature within ECOR or GMB strains, or within phylogroups, we used different groups to compute the dN/dS ratio (**Table 3.2**).

Table 3.2.dN/dS ratio for 8 loci in different groups of isolates. Calculations performed with the START2 package. The phylogroups groupings used here are composed of ECOR and GMB strains indiscriminately.

	ECOR (n=66)	GMB (n=76)	All (n=142)	A (n=37)	B1 (n=49)	B2 (n=21)	D (n=25)	E (n=7)
dinB	0.026	0.026	0.026	0.022	0.044	0.014	0.035	0.058
icdA	0.032	0.007	0.022	0.031	0.046	0.012	0.033	0.009
pabB	0.102	0.118	0.116	0.192	0.186	0.110	0.138	0.373
polB	0.049	0.029	0.041	0.049	0.050	0.124	0.007	0.068
putP	0.057	0.020	0.043	0.029	0.081	0.073	0.056	0.011
trpA	0.024	0.033	0.029	0.000	0.034	0.084	0.033	0.009
trpB	0.031	0.021	0.030	0.324	0.078	0.000	0.009	0.028
uidA	0.049	0.042	0.048	0.095	0.044	0.100	0.072	0.055

Even within different groups of strains, there was a higher dN/dS ratio at the *pabB* locus, especially in phylogroup E strains (**Table 3.2**). Overall, this higher trend, possibly caused by a slight ongoing positive selection at this locus, was constant in all phylogroups. Interestingly, we could detect possible traces of ongoing positive selection in *trpB* encoding the tryptophan synthase subunit B, but only in strains from phylogroup A. This observation highlights the importance of checking the nature of selection, as sub-groups in a tested population can exhibit varying selection pressures signatures at the same loci. Overall, all tested loci from this MLST scheme were under strong purifying selection, which is a common feature for housekeeping genes (Perez-Losada et al., 2007). From this analysis, we can conclude that the *E. coli* MLST scheme that we used targets genes that are likely to be good indicators of the phylogenetic signal, exempt of a strong positive selection effect.

3.4.1.3. Clonal relationships as examined by MLST

For each gene analysed by MLST, we obtained different “alleles” (or variants of the same gene) with unique sequences. One allelic sequence at one locus is assigned a number, or “allele-type” (AT) and the combination of 8 ATs is called the “sequence type” (ST). Two MLST clones are defined as isolates with the same ST, meaning that the 8 sequences at the different tested loci (i.e. the 8 ATs) are identical. Two isolates can also be very closely related but not clones, as for instance they can share 6 or 7 out of 8 ATs. Such clones are often grouped in related clonal complexes (Spratt, 2004). As the principle of MLST is to compare the sequences of multiple loci that are thought not to be under strong positive selection and thus accumulate variation much more slowly, even if one particular allele comes from homologous recombination from a completely unrelated isolate, it is very unlikely that the 6 or 7 other genes are similar to that isolate too, and the real phylogenetic relationships are preserved. One has then to be very cautious when interpreting phylogenies based on variation at a single locus, or at genes subjected to high levels of recombination or mutation (typically those under positive selection, e.g. surface proteins that are in direct contact with their host immune system) as these phylogenies presumably do not take into account recombination the way MLST does and can convey false information on the relationships between isolates because of it (Spratt, 2004).

In this study, we used MLST to compare 76 GMB and 66 ECOR strains. The scheme we used involves the sequencing of internal fragments of 8 housekeeping genes for a concatenated total of 4,095 bp (Jaureguy et al., 2008). The Pasteur MLST database was used to assign our GMB isolates to already existing ATs and STs and determine

which were new. Information used in this work (ST and AT numbers already present in the database) is based on the 20/07/2011 update of the MLST database. For new ATs and STs, a number was assigned in this work, before submission of our sequences to the database curator, so it may not correspond to the final number appearing in the most up to date database version (after 20/07/2011). At the time of our study, there were 104 AT_{dinB}, 202 AT_{icdA}, 127 AT_{pabB}, 157 AT_{putP}, 138 AT_{trpA}, 137 AT_{trpB}, 128 AT_{uidA} for a total of 522 unique STs. Using our GMB strains, we defined new ATs with a number starting from 300, and new STs starting with 600 (**Table 3.3**).

Table 3.3.Allelic profiles of ECOR and GMB strains tested by MLST (see text for more details). The list below is ordered by ST number.

Isolate	ST[a]	Year[b]	Isolated from[b]	Location	AT[a]							
					dinB	*icd A*	*pab B*	*pol B*	*put P*	*trp A*	*trp B*	*uidA*
ECOR61	1	1980s	Human (Female)	Sweden	1	1	2	1	1	2	3	1
ECOR62	1	1980s	Human (Female)	Sweden	1	1	2	1	1	2	3	1
ECOR10	2	1980s	Human (Female)	Sweden	8	2	7	3	7	1	4	2
GMB90	2	2009	Spinach	Norfolk, UK	8	2	7	3	7	1	4	2
GMB15	3	2008	Organic spinach	Norfolk, UK	3	8	5	11	8	3	5	3
GMB50	3	2008	Organic spinach	Berkshire, UK	3	8	5	11	8	3	5	3
GMB103	21	2009	Babycorn	Outside UK	7	33	18	2	5	28	2	2
GMB22	21	2008	Teen spinach	Dorset, UK	7	33	18	2	5	28	2	2
GMB38	21	2008	Baby spinach	Berkshire, UK	7	33	18	2	5	28	2	2
GMB21	48	2008	Baby spinach	Norfolk, UK	2	11	23	15	10	15	10	12
GMB30	48	2008	Baby spinach	Dover, UK	2	11	23	15	10	15	10	12
ECOR56	52	1980s	Human (Female)	Sweden	2	4	6	4	1	21	1	1
GMB24	66	2008	Organic spinach	Norfolk, UK	6	5	3	2	6	7	2	4
GMB104	83	2009	Spinach	Norfolk, UK	11	3	4	3	15	1	4	16
GMB43	86	2008	Baby spinach	Dorset, UK	24	31	4	26	16	29	2	2
GMB81	86	2008	Rocket	Norfolk, UK	24	31	4	26	16	29	2	2
GMB17	108	2008	Mizuna	Norfolk, UK	25	65	48	10	16	8	2	2
GMB10	117	2008	Organic spinach	Norfolk, UK	5	47	3	10	6	7	4	2
GMB45	122	2008	Organic spinach	Various	2	73	2	55	43	18	46	1
ECOR18	132	1980s	Celebese ape	USA (Wash.)	10	2	7	3	7	1	4	2
ECOR01	163	1980s	Human (Female, 19yr)	USA (Iowa)	8	85	7	3	58	1	57	2
ECOR11	164	1980s	Human (Female)	Sweden	8	2	7	3	59	1	4	2
ECOR12	165	1980s	Human (Female)	Sweden	8	86	7	3	7	1	4	2
ECOR14	166	1980s	Human (Female)	Sweden	57	2	7	3	60	1	58	23
ECOR15	167	1980s	Human (Female)	Sweden	10	87	7	3	18	1	57	2
ECOR16	168	1980s	Leopard	USA (Wash.)	10	2	7	17	18	1	57	67
ECOR17	169	1980s	Pig	Indonesia	10	28	7	3	18	1	57	23
ECOR19	170	1980s	Celebese ape	USA (Wash.)	10	2	7	3	61	1	57	2

ECOR02	171	1979	Human (Male)	USA (N.Y.)	8	2	7	66	59	1	4	2
ECOR20	172	1980s	Steer	Bali	10	88	7	3	7	1	59	2
ECOR21	173	1980s	Steer	Bali	10	29	7	3	7	1	60	2
ECOR22	174	1980s	Steer	Bali	10	28	7	67	18	1	4	23
ECOR23	175	1980s	Elephant	USA (Wash.)	10	89	64	3	18	1	4	23
ECOR24	176	1980s	Human (Female)	Sweden	58	28	3	17	16	1	57	4
ECOR25	177	1980s	Dog	USA (N.Y.)	8	90	7	3	7	1	4	2
ECOR26	178	1980s	Human (infant)	USA (Mass.)	25	91	3	10	62	29	61	2
ECOR27	179	1980s	Giraffe	USA (Wash.)	25	92	3	10	26	29	62	2
ECOR28	180	1980s	Human (Female, 4yr)	USA (Iowa)	25	93	3	10	6	64	57	68
ECOR29	181	1980s	Kangaroo rat	USA (Nev.)	25	47	48	68	5	29	63	2
ECOR03	182	1980s	Dog	USA (Mass.)	8	2	7	3	7	65	4	2
ECOR30	183	1980s	Bison	Canada	25	47	48	68	5	29	64	2
ECOR32	184	1980s	Giraffe	USA (Wash.)	5	47	48	68	63	29	63	2
ECOR33	185	1980s	Sheep	USA (Calif.)	5	47	48	68	5	29	63	2
ECOR34	186	1980s	Dog	USA (Mass.)	7	94	65	68	63	8	2	2
ECOR37	187	1980s	Marmoset	USA (Wash.)	59	95	66	63	64	66	65	69
ECOR04	188	1980s	Human (Female, 5yr)	USA (Iowa)	10	2	3	69	18	1	57	23
ECOR41	189	1982	Human (Female, 22yr)	Tonga	31	46	17	12	12	67	26	70
ECOR42	190	1979	Human (Male)	USA (Mass.)	35	42	67	70	65	36	66	71
ECOR43	191	1980s	Human (Female)	Sweden	8	2	7	3	61	1	4	2
ECOR44	192	1980s	Cougar	USA (Wash.)	17	96	68	71	66	12	13	32
ECOR45	193	1980s	Pig	Indonesia	60	97	4	10	67	7	4	2
ECOR46	194	1980s	Ape	USA (Wash.)	18	12	17	14	68	68	67	14
ECOR50	195	1980s	Human (Female)	Sweden	17	9	28	3	9	13	68	72
ECOR51	196	1980s	Human (infant)	USA (Mass.)	2	4	69	72	69	6	69	1
ECOR52	197	1980s	Orangutan	USA (Wash.)	2	46	6	4	1	6	69	1
ECOR53	198	1980s	Human (Female, 4yr)	USA (Iowa)	4	19	1	73	2	2	1	1
ECOR54	199	1980s	Human	USA (Iowa)	2	98	70	74	1	6	69	1
ECOR55	200	1980s	Human	Sweden	2	4	6	74	1	6	1	1
ECOR57	201	1980s	Gorilla	USA (Wash.)	2	4	6	74	70	6	1	1
ECOR58	202	1980s	Lion	USA (Wash.)	2	3	4	68	71	57	4	2

GMB59	202	2008	Rocket	Norfolk, UK	2	3	4	68	71	57	4	2
ECOR59	203	1979	Human (Male)	USA (Mass.)	4	19	1	6	72	2	1	1
ECOR60	204	1980s	Human (Female)	Sweden	4	19	71	6	73	2	1	1
ECOR63	205	1980s	Human (Female)	Sweden	2	15	72	75	2	20	1	73
ECOR65	206	1980s	Celebese ape	USA (Wash.)	1	23	73	55	74	18	46	1
ECOR66	207	1980s	Celebese ape	USA (Wash.)	16	99	74	76	75	9	6	74
ECOR67	208	1980s	Goat	Indonesia	16	100	3	76	76	69	62	74
ECOR68	209	1980s	Giraffe	USA (Wash.)	25	101	4	77	77	53	62	2
ECOR07	210	1980s	Orangutan	USA (Wash.)	10	2	7	3	61	1	4	2
ECOR70	211	1980s	Gorilla	USA (Wash.)	6	5	75	2	6	28	62	4
ECOR71	212	1980s	Human (Female)	Sweden	25	65	3	78	78	28	62	2
ECOR72	213	1980s	Human (Female)	Sweden	25	34	4	23	6	7	62	23
ECOR08	214	1980s	Human (Female, 20yr)	USA (Iowa)	8	2	7	79	59	1	57	2
ECOR31	250	1980s	Leopard	USA (Wash.)	22	41	21	3	23	87	89	18
ECOR35	251	1980s	Human (Female, 36yr)	USA (Iowa)	31	113	24	36	98	24	90	70
ECOR38	253	1980s	Human (Female, 21yr)	USA (Iowa)	31	46	24	36	12	67	90	70
ECOR40	253	1980s	Human	Sweden	31	46	24	36	12	67	90	70
ECOR39	254	1980s	Human	Sweden	31	46	24	36	12	67	26	70
ECOR47	255	1980s	Sheep	New Guinea	59	114	89	11	99	3	5	3
ECOR48	256	1980s	Human (Female)	Sweden	64	10	90	85	100	88	91	11
ECOR49	257	1980s	Human (Female)	Sweden	17	115	28	12	101	13	9	84
ECOR69	258	1980s	Celebese ape	USA (Wash.)	24	31	4	26	50	29	62	2
GMB07	303	2008	Spinach	Dover, UK	25	3	48	10	26	57	4	90
GMB47	319	2008	Spinach	Dover, UK	73	135	102	76	121	72	1	95
GMB14	338	2008	Mizuna	King's Lynn, UK	24	3	3	26	16	108	4	2
GMB40	352	2008	Baby spinach	Dover, UK	2	23	73	55	43	18	46	1
GMB76	352	2008	Rocket	Norfolk, UK	2	23	73	55	43	18	46	1
GMB89	363	2009	Spinach	Norfolk, UK	50	65	3	2	5	111	2	2
GMB101	366	2009	Organic seasonal	Various	5	47	48	68	5	29	109	2
GMB48	446	2008	Baby spinach	Norfolk, UK	10	2	3	3	7	1	4	2
GMB102	512	2009	Rocket	Dorset, UK	25	37	4	10	84	7	4	2
GMB41	600	2008	Tatsoi	Dorset, UK	202	33	18	2	5	8	2	2

GMB65	601	2008	Rocket	Norfolk, UK	104	3	3	10	5	8	2	2
GMB107	602	2009	Salad mix bag	Various	200	37	4	10	78	8	2	2
GMB37	603	2008	Baby spinach	Dover, UK	8	2	3	3	7	1	4	2
GMB20	604	2008	Wild rocket	Norfolk, UK	8	28	7	3	7	1	4	2
GMB108	605	2009	Spinach	Norfolk, UK	8	201	7	3	7	1	4	2
GMB23	606	2008	Teen spinach	Dover, UK	8	204	7	3	7	1	4	2
GMB35	607	2008	Red Chard	Norfolk, UK	8	205	7	3	7	1	4	2
GMB01	608	2008	Wild rocket	Outside UK	10	2	3	52	7	1	4	2
GMB02	609	2008	Wild rocket	Outside UK	10	2	7	83	7	1	4	2
GMB06	610	2008	Organic spinach	Berkshire, UK	11	196	200	17	18	1	4	2
GMB58	611	2008	Red Amaranth	Various	204	209	3	10	78	1	4	2
GMB03	612	2008	Wild rocket	King's Lynn, UK	60	200	4	10	67	7	4	2
GMB25	613	2008	Spinach	Norfolk, UK	7	47	3	52	16	57	4	2
GMB05	614	2008	Mizuna	King's Lynn, UK	24	3	3	26	16	200	4	2
GMB32	615	2008	Spinach	Dorset, UK	10	2	3	3	7	1	201	2
GMB33	615	2008	Spinach	Dorset, UK	10	2	3	3	7	1	201	2
GMB34	615	2008	Spinach	Dorset, UK	10	2	3	3	7	1	201	2
GMB92	616	2009	Rocket	Dorset, UK	5	47	4	10	26	1	206	2
GMB100	617	2009	Spinach	Various	1	13	2	31	1	4	1	6
GMB73	618	2008	Rocket	Norfolk, UK	51	13	204	4	21	2	3	6
GMB04	619	2008	Wild rocket	King's Lynn, UK	18	8	112	11	8	12	13	11
GMB12	619	2008	Mizuna	Berkshire, UK	18	8	112	11	8	12	13	11
GMB44	620	2008	Salad mix bag	Various	2	206	23	15	10	15	10	12
GMB84	621	2008	Soil	Norfolk, UK	10	148	3	3	18	1	4	23
GMB91	621	2009	Spinach baby leaf	Various	10	148	3	3	18	1	4	23
GMB28	622	2008	Organic spinach	Berkshire, UK	5	65	3	10	16	8	2	30
GMB80	622	2008	Rocket	Norfolk, UK	5	65	3	10	16	8	2	30
GMB54	623	2008	Spinach	Norfolk, UK	18	208	112	94	66	12	13	32
GMB77	624	2008	Rocket	Norfolk, UK	22	64	205	202	202	37	205	46
GMB78	624	2008	Rocket	Norfolk, UK	22	64	205	202	202	37	205	46
GMB79	625	2008	Rocket	Norfolk, UK	25	210	3	203	16	57	4	50
GMB29	626	2008	Rocket and chard	Norfolk, UK	50	47	3	10	5	7	4	55
GMB88	627	2009	Spinach	Norfolk, UK	25	3	3	10	78	1	16	57

GMB16	628	2008	Mizuna	Norfolk, UK	5	47	4	56	6	7	4	58
GMB53	629	2008	Organic spinach	Norfolk, UK	25	47	48	10	26	57	4	113
GMB18	630	2008	Spinach	Dorset, UK	5	202	4	56	6	29	2	200
GMB19	631	2008	Tatsoi	Dorset, UK	17	203	201	11	8	201	200	201
GMB39	632	2008	Baby spinach	Dorset, UK	201	67	202	39	200	202	202	202
GMB46	633	2008	Organic spinach	Various	203	207	203	200	201	203	203	203
GMB56	633	2008	Spinach mix	Various	203	207	203	200	201	203	203	203
GMB57	633	2008	Spinach mix	Various	203	207	203	200	201	203	203	203
GMB60	634	2008	Rocket	Norfolk, UK	205	74	58	15	2	204	10	204
GMB61	635	2008	Rocket	Norfolk, UK	206	10	90	201	30	16	204	205
GMB63	635	2008	Rocket	Norfolk, UK	206	10	90	201	30	16	204	205
GMB66	635	2008	Rocket	Norfolk, UK	206	10	90	201	30	16	204	205
GMB74	635	2008	Rocket	Norfolk, UK	206	10	90	201	30	16	204	205
GMB64	636	2008	Rocket	Norfolk, UK	25	3	4	68	71	1	16	206
GMB83	637	2008	Soil	Norfolk, UK	25	3	48	10	26	57	4	207
GMB93	638	2009	Spinach	Norfolk, UK	88	15	127	76	4	205	1	208
GMB98	639	2009	Baby spinach	Dorset, UK	4	13	9	29	1	32	1	209

[a] ST numbers below 600 (and AT below 200) correspond to ST (AT) present in the database after its last update at the time of the study (20/07/2011); ST numbers above 600 (and AT above 200) were arbitrarily attributed for this study only and may not reflect what is present on the database after its 20/07/2011 update.

[b]For ECOR strains, information is as indicated on the ECOR website (http://foodsafe.msu.edu/whittam/ecor/; last accessed 28/09/2011)

In our MLST analysis of 76 GMB strains, we identified 7 new AT_{dinB}, 11 new AT_{icdA}, 6 new AT_{pabB}, 4 new AT_{polB}, 3 new AT_{putP}, 6 new AT_{trpA}, 7 new AT_{trpB} and 10 new AT_{uidA} (arbitrarily labelled from AT number 200 in **Table 3.3**). The analysis of GMB strains defined 40 new STs that were not described in the Pasteur database (arbitrarily labelled 600 to 639 in **Table 3.3**), both from ATs already present in the database but in a unique combination, and from new ATs. Among GMB strains, 34/76 (44.7%) had at least one new AT (not present in the database) and 42/76 (55.3%) were already fully defined in the MLST database.

To examine if plant isolates could be related to representative strains of the *E. coli* species as symbolised by ECOR strains, we compared sequences of 76 GMB isolates with those of 66 ECOR strains already present in the Pasteur MLST database (**Table 3.3**). In total, 15 STs had more than one strain assigned to them. We inferred that GMB strains sharing the same ST and isolated at the same time and location were likely to be very recent clones, corresponding to the same strain isolated at the same time (**Table 3.4**).

Table 3.4.**Clones likely to be very recent among strains tested by MLST**. Recentness was deduced for clones that were isolated at the same time and location.

Isolate	ST	Year	Source	Location
GMB32	615	2008	Spinach	Dorset, UK
GMB33				
GMB34				
GMB46	633	2008	Spinach	Various
GMB56				
GMB57				
GMB61	635	2008	Rocket	Norfolk, UK
GMB63				
GMB66				
GMB74				
GMB77	624	2008	Rocket	Norfolk, UK
GMB78				

Only 2 STs (ST-2 and ST-202) regrouped GMB and ECOR strains. Also, there were seemingly unrelated GMB strains sharing the same ST (**Table 3.5**) and in 2 occurrences, strains isolated in 2008 and 2009 shared the same ST (**Table 3.5**).

Table 3.5.**Clones within tested GMB strains with no apparent isolation links.**

Isolate	ST	Year	Source	Location
GMB15	3	2008	Spinach	Norfolk, UK
GMB50				Berkshire, UK
GMB103	21	2009	Babycorn	Outside UK
GMB22		2008	Spinach	Dorset, UK
GMB38				Berkshire, UK
GMB21	48	2008	Spinach	Norfolk, UK
GMB30				Dover, UK
GMB43	86	2008	Spinach	Dorset, UK
GMB81			Rocket	Norfolk, UK
GMB40	352	2008	Spinach	Dover, UK
GMB76			Rocket	Norfolk, UK
GMB04	619	2008	Rocket	King's Lynn, UK
GMB12			Mizuna	Berkshire, UK
GMB84	621	2008	Soil	Norfolk, UK
GMB91		2009	Spinach	Various
GMB28	622	2008	Spinach	Berkshire, UK
GMB80			Rocket	Norfolk, UK

Together, these observations suggest that some clones are probably ubiquitous and conserved in *E. coli*, such as ST-2 and ST-202 which are both regrouping ECOR and GMB strains isolated 30 years apart on different continents (ST-202). Other GMB clones can be isolated from one year to another (2008 and 2009) and simultaneously from soil and plant. This last point is interesting as it would suggest that there is persistence of *E. coli* in soils and transfer from soil to plants. However, there are not enough soil samples in the GMB collection to make clear assumptions, and the two isolates in ST-621 (GMB84 from soil in 2008 and GMB91 from spinach in 2009) were not isolated in the same location.

We performed a complementary analysis of clonality by comparing the BOX profiles of MLST clones (**Figure 3.11**). MLST focuses on a limited number of neutrally-evolving housekeeping genes, which gives a powerful signal for phylogenetic history inference, but does not take into account the other parts in the genome that may vary, and BOX-PCR may be a good way to capture that variability.

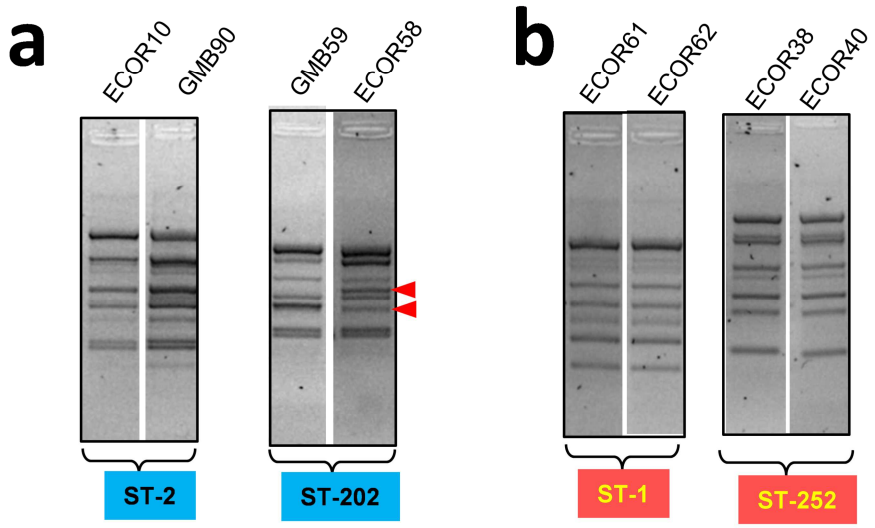


Figure 3.11. **BOX-PCR electrophoretic profiles of MLST clones**. (a) clones from both GMB and ECOR collections and (b) ECOR only clones. Red arrows indicate varying bands on the profiles.

Profiles of strains from the same ST were most of the time very similar, with a few differences in the number of bands or intensity in some cases (**Figure 3.11a** and **3.12a**). Profiles of strains within ST-2 or ST-202 (containing both GMB and ECOR strains) were remarkably similar, once again suggesting that despite being isolated 30 years apart at different geographical locations, some *E. coli* strains remain fairly identical. Similarly, ECOR clones also had identical BOX profiles (**Figure 3.11b**).

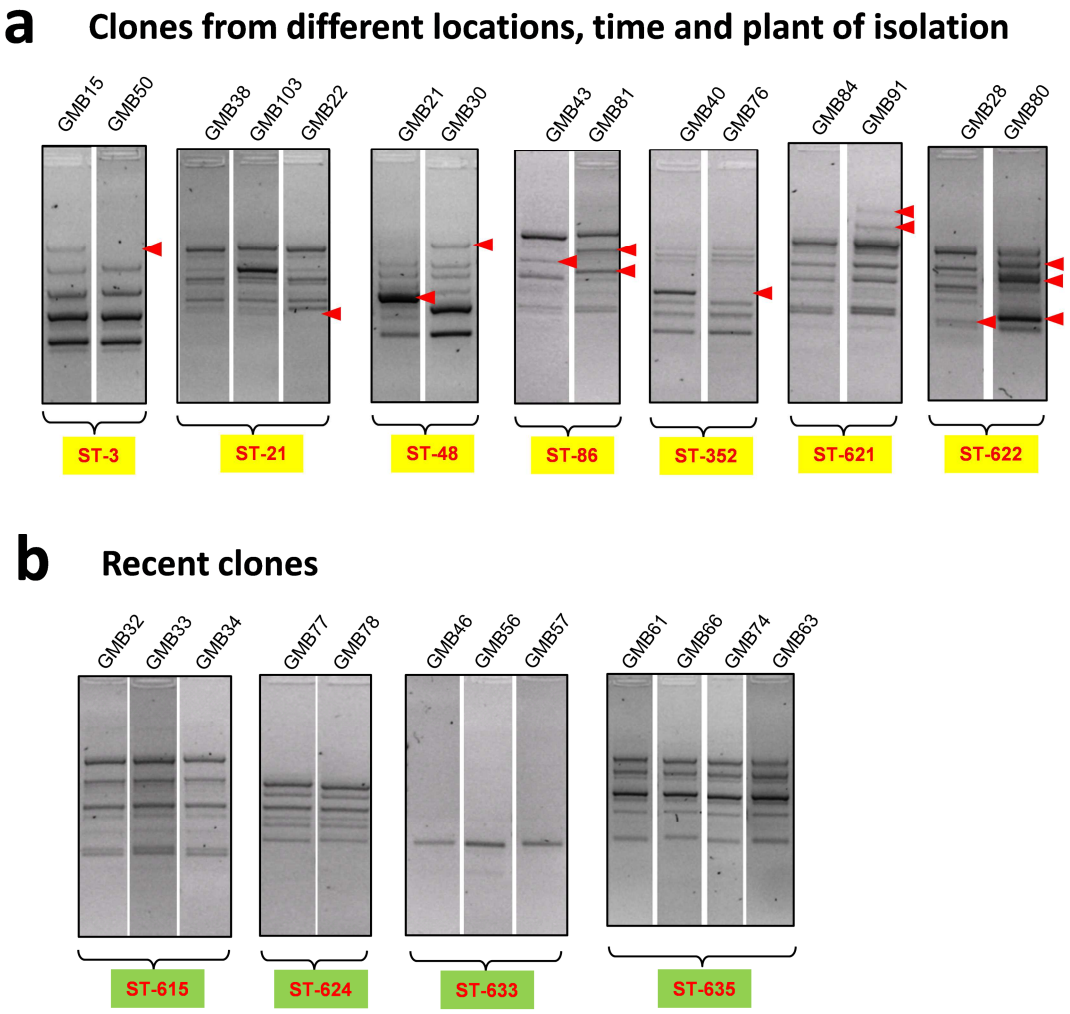


Figure 3.12. BOX-PCR electrophoretic profiles of GMB clones. (a) clones isolated at different times, locations and plants (b) GMB clones likely to be recent (same time and location of isolation). Red arrows indicate varying bands on the profiles.

Interestingly, there were no differences at all in BOX profiles for GMB clones that are believed to be recent (**Figure 3.12b**) but there were slight differences for GMB clones that are less related in time and location from each other. It is not surprising to see that isolates that are phylogenetically very close but isolated at different times and locations, have experienced genomic rearrangements that are not detected using MLST. On the other hand, recent clones are believed to have emerged from the same

ancestor cell, and in our case maybe even on the plant it was isolated from. It is then not surprising as well to observe that these strains share identical BOX profiles.

The fact that these differences are not detected by MLST does not hinder the power of it. Two MLST clones, even distantly isolated in time and location are very likely to be similar as they are not observed to be phylogenetically divergent. Evolutionary forces that generate variability at neutral loci such as the ones observed with MLST did not have time or ecological opportunity to exert on two clones from the same ST. Some genomic rearrangements can obviously occur independently of the phylogenetic history as captured by MLST (acquisition of foreign DNA, prophage insertions, genomic rearrangements). This additional information (unrelated to phylogenetic history) can be observed just as we did, using DNA fingerprinting methods, or by sequencing and comparing whole genomes.

3.4.1.4. Diversity estimators

It is also possible to numerically represent diversity within a population or collection of samples via diversity estimators. Based on allelic profiles of tested isolates, we used rarefaction analysis, calculated the Chao1 estimator and the abundance coverage estimator (ACE) which are both considered to be the least biased after comparison studies with other ways of estimating richness (see Methods section 2.4.3).

We first computed rarefaction curves as an indirect method to observe diversity in *E. coli* environmental isolates (**Figure 3.13**). In a recent review (Hughes and Hellmann, 2005), a rather clear analogy was mentioned to explain the concept of a rarefaction

curve. If you consider a bird-spotter walking in a forest and recording the species of encountered spotted birds, data can be summarised by a classical "accumulation curve", corresponding to the plot of the cumulative number of bird species observed as each bird species is spotted and recorded in the forest environment. Classical accumulation curves are not very meaningful for microbiologists as, contrary to bird-spotters, the geographical transect along which bacterial species are detected in a given environment does not usually carry as much ecological sense as it does for higher organisms. This is not true for rarefaction curves, which are the estimated smoothed average of all possible accumulation curves and represent the average number of bird species observed when individuals are drawn with replacement from the same pool of individuals (or "sample") over and over. Recent and common uses include the estimations of species richness based on 16S rDNA sequencing samples (Qin et al., 2010; Quaiser et al., 2011). The method assumes that there is a finite number of distinct species (or any other taxonomic unit) in a given environment and that experimental sampling only gives an incomplete representation of the total richness. Thus, more sampling is likely to uncover more distinct species, but as the species number in a given environment is finite, it will take increasingly more sampling effort to uncover new distinct species.

In this study, we used rarefaction curves to estimate the required sampling effort to reach the estimated maximum number of possible STs for *E. coli*. We compared rarefaction curves for GMB and ECOR strains, by assuming that a curve closer to the proportionality line reflects a higher diversity. ECOR is expected to have the highest diversity, as the collection was assembled *de novo* from a much larger sampling of isolates with the precise purpose of representing the highest possible diversity in *E.*

coli. In that retrospect, ECOR is not a coherent ecological population (i.e., all strains were isolated in the same conditions or hosts), but it is nevertheless a good substitute for approximating the whole *E. coli* species diversity, and this is why we included it in our diversity analyses.

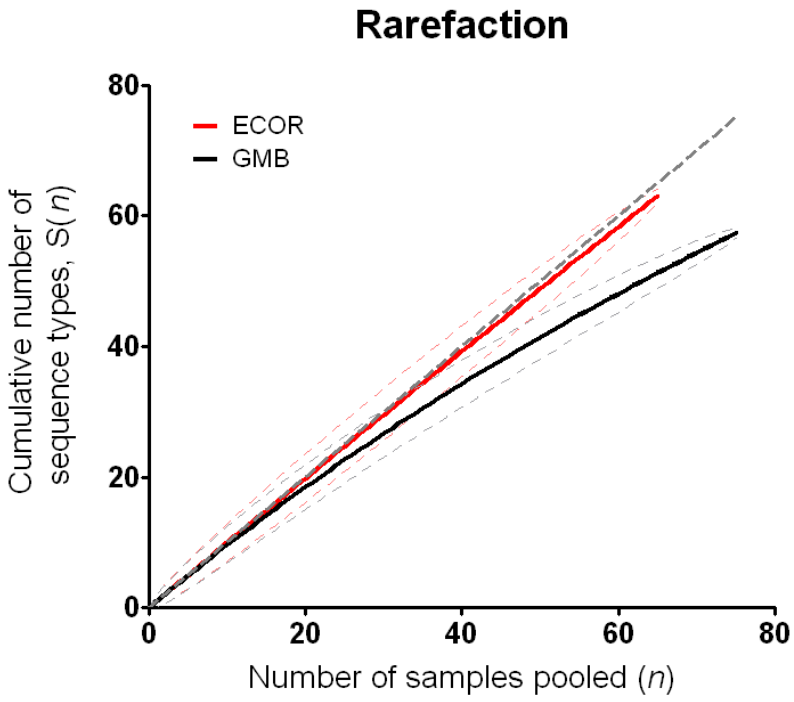


Figure 3.13. Rarefaction analysis of two *E. coli* collections. Red line correspond to ECOR with data in the Pasteur MLST database, black line correspond to all GMB strains included in the MLST analysis.

ECOR had a higher diversity than GMB (**Figure 3.13**). As the 72 ECOR strains were precisely selected from a larger collection to encompass the maximum genetic diversity within the *E. coli* species, the observation that any punctually sampled *E. coli* strains are less diverse is not surprising.

We then estimated the maximum sampling effort required to capture all diversity in *E. coli*, should evolutionary constraints sampled environments remain identical. Based on the EstimateS calculation of rarefaction curves, we fitted polynomial curves of known equations and calculated the maximum *y* values for each rarefaction curves to obtain the theoretical highest number of STs for each sampling environment (**Figure 3.14**).

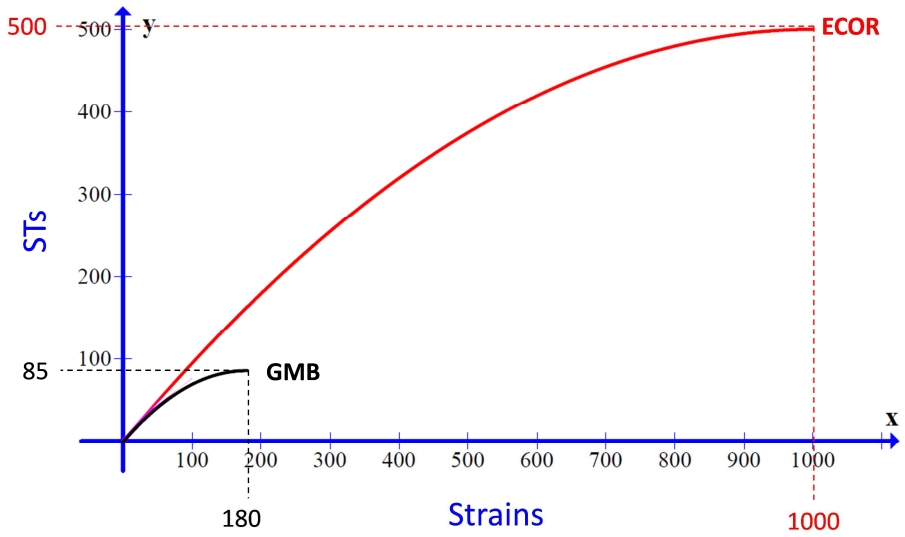


Figure 3.14. Estimation of the theoretical sampling effort to capture all diversity in *E. coli* GMB and ECOR collections. Red line corresponds to ECOR with data in the Pasteur MLST database, black line corresponds to all GMB strains included in the MLST analysis.

The sampling effort required to capture all diversity was expectedly very high for ECOR (1000 strains of 500 different STs). Obviously, the calculation reflects the synthetic nature of the ECOR collection. More relevantly, the sampling effort required to capture the whole diversity of *E. coli* on plants (based on GMB diversity) was found to be 180 strains for 85 STs. Our current sampling then represents 37.2% of the predicted sample size that would approximately capture all diversity of *E. coli* from plants.

It is important to keep in mind that the notion of sampling effort is a theoretical way to compare diversity and is not likely to carry a strong biological meaning. Sampling effort approximation is calculated from the observed collection, isolated under precise conditions and using a specific MLST scheme. In order for this calculation to be correct, the constraints on the sampling environment would have to remain identical for the whole additional sampling effort, which is unlikely. Also, rarefaction has been extrapolated to MLST studies from species diversity studies (de Muinck et al., 2011), which arguably behave differently than same-species individuals.

Another commonly accepted way of estimating diversity is through the calculations of diversity estimators. Similarly to rarefaction analysis, these calculations are usually used to examine bacterial diversity in a mixed sample, but can also be extrapolated to the analysis of MLST data. The output of our calculations is also presented in numbers of strains required to capture the maximum diversity in a given sample. Using AT and ST definitions and distribution in GMB and ECOR strains, we calculated and compared the Chao1 and ACE indices of diversity (**Figure 3.15**), both described as least biased estimators in comparative statistical studies (See Methods section 2.4.3), and also for comparison with previous studies (Walk et al., 2007; de Muinck et al., 2011). These estimators are calculated based on the number of tested strains, and our method of calculation presented here (using the EstimateS software method) evaluates and adjusts average richness estimators for all steps of species accumulation, and therefore produces a curve as output (**Figure 3.15**).

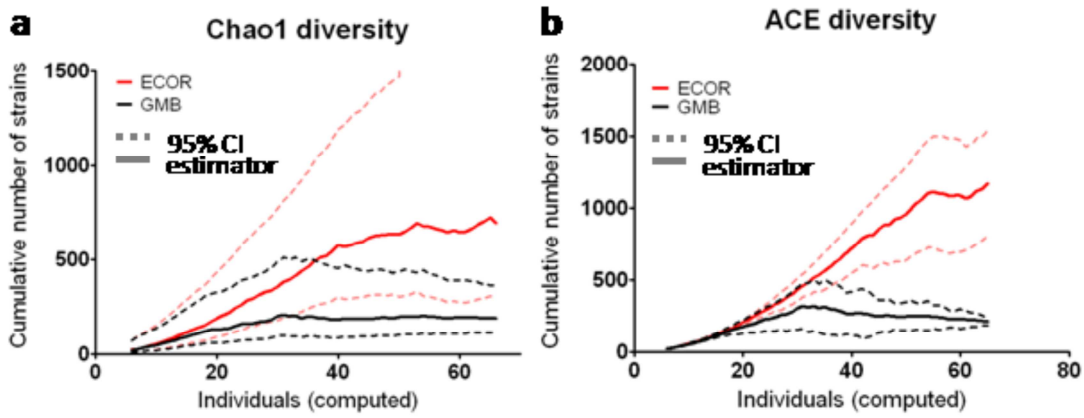


Figure 3.15. ECOR and GMB strain diversity using Chao1 and ACE estimators. Plain lines refer to the estimator (Chao1 or ACE) and dotted lines refer to the 95% confidence interval (CI) associated to the estimator calculation. Calculations were made using the EstimateS v7.5 software.

Overall, there was not much numerical difference between the two estimators. ACE is described to be more accurate but we included Chao1 as it is one of the most common estimator calculated in diversity studies. As observed with rarefaction and as expected, the diversity of GMB strains was lower than ECOR. GMB strain diversity seems to reach a constant top value around 180 to 250 strains using both estimators (**Figure 3.15**), very similarly to what was found with the sampling effort analysis (**Figure 3.14**).

Even if the concept of sampling effort yield approximate results, it is useful to compare calculations across studies. Previous work on *E. coli* faecal isolates and their vertical transmission after birth identified that transmitted strains were much less diverse in infants than in their mothers, but that diversity levels increased as the children aged (de Muinck et al., 2011). Overall, Chao1 and ACE diversity was much lower in this study than in our GMB strains. This difference could be artificial (diversity was calculated using one gene only) or biological, as it has been observed

that *E. coli* diversity was very low when sampled from individual hosts, in which only one or two single clones are dominant (Wallick and Stuart, 1943; Sears et al., 1950; Sears and Brownlee, 1952; Cooke et al., 1972; Smith, 1975). One can then imagine that it would take a large sample of numerous individuals to cover a large diversity of isolates, which is not the case in the study mentioned above (de Muinck et al., 2011). Another study on isolates from freshwaters found a Chao1 diversity average ranging from 200 to 400 STs between different sampling sites (Walk et al., 2007), which is slightly more than our calculations for GMB diversity (**Figures 3.14** and **3.15**). From this comparison, we can suggest that *E. coli* diversity is observed to be much greater when nonhost secondary environments are sampled (freshwaters, plants) than when faecal isolates from mammalian hosts are examined.

3.4.2. Reconstruction of phylogenetic relationships between GMB and ECOR strains

It is possible to infer relationships between strains at a deeper phylogenetic level using the output of MLST. Either allele-based or sequence-based information provide information at different levels. Allele-based approaches such as eBURST and minimum spanning trees (MS_{TREE}) analyses are useful to examine relatedness of isolates in the light of gene flow and allelic exchange or recombination. Sequence-based approaches examine single nucleotide polymorphisms and substitutions across a set of isolates to provide finer phylogenetic reconstructions.

3.4.2.1. Allele-based population genetic structure of GMB and ECOR strains

Using allelic information (ST numbers and their corresponding defining ATs), it is possible to examine relatedness between isolates (or STs) and the population structure of microbial species with intermediate levels of recombination, such as *E. coli* (Wirth et al., 2006). Different STs can be closely related if they share most of their ATs or relatively distant if none of their ATs are the same. This could be simply explained by gene flow, or recombination events at particular alleles. Conversely, 2 divergent clones can share most of their AT but differ at one locus where mutations have accumulated and constitute distinct ATs at this locus.

Simple algorithms have been developed to examine and relate STs based on how many AT they share. Typically, when 2 STs share *n*–2 alleles (where *n* is the total number of alleles defining the ST; in our case, *n*=8 and *n*–2=6), they are defined to be in the same clonal complex. In this work, we used the PhyloViz software (http://www.phyloviz.net/wiki/) to construct minimum spanning trees, or MS_{TREE} based on a BURST (for "Based Upon Related Sequence Types") calculation (**Figure 3.17**). Using the BURST algorithm, ancestral STs were defined, and variants are mapped around it on different levels with single locus variants (SLV), corresponding to STs sharing *n*–1 ATs, closer than double loci variants (DLV), corresponding to STs sharing *n*–2 ATs.

There were 4 clonal complexes in our dataset of ECOR and GMB strains with ST-2, ST-132, ST-185 and ST-446 as ancestral STs (**Figure 3.17**). These clonal complexes

are representative of how diversity arises in single-species populations in the sense that some allelic profiles are shown to be more stable over time and selective pressures than others. In our study, this is for instance the case for ST-2 from which 11 other STs are related at *n*–1 and *n*–2. It is possible that strains from ST-2 possess traits that confer a broad ecological stability in the *E. coli* species, as both ECOR and GMB strains, isolated 30 years and continents apart, are from ST-2 or related.

MS_{TREES} are calculated using a model assuming that allelic profiles evolution is explained with as few events as possible. The obtained tree is therefore representing the simplest, shortest combination of allelic profiles changes between all tested allelic profiles. This simplicity results in the fact that contrary to other models of phylogenetic inference, MS_{TREE} calculation does not hypothesise putative internal nodes of expected common ancestors. Therefore, all samples are linked together according to their similarity, regardless of their true phylogenetic history. Because of these limitations, one has to provide all the possible intermediate samples, which must not have a lot of variation between them in order to be meaningfully linked. The simpler analysis also has advantages, in the sense that MS_{TREES} are focused primarily on micro-evolution and short-term divergence, giving more power to variation forces that can be considered low using evolution-based models. Also, when coupled with a BURST analysis, MS_{TREES} are a way of mapping the simplest possible links between each clonal complex, therefore providing a simple and informative insight into how all complexes and STs are related to each other. MS_{TREES} are commonly used in epidemiology studies based on MLST, and are often seen as a simple and preliminary phylogenetic analysis method (Wirth et al., 2006; Millet et al., 2009; Bunnik et al., 2011; Mellmann et al., 2011). Here, we present the same MS_{TREE} coloured by 5

different parameters: collection (ECOR or GMB), source, location and year of isolation, and phylogroup (**Figure 3.17**).

→**Figure 3.17** (5 next pages). **Minimum spanning tree (MS_{TREE})of BURST outputs based on 142 allelic profiles after MLST**. Different colouring were applied to the same MST: (a) collection (ECOR or GMB); (b) source of isolation (human, primate, non-primate, spinach, rocket, other salad and soil); (c) year of isolation (1980s, 2008 and 2009); (d) location of isolation (UK, Sweden, Italy, USA, Canada, Asia and Pacific); (e) phylogroups (A, B1, B2, D, E, outgroup). The thickness of the link between two nodes reflects the relatedness of the corresponding STs with black for SLV or DLV and grey for links at n>2 shared allele types. The figures were obtained using the PhyloViz software (http://www.phyloviz.net/wiki/). See text for more details.

Figure 3.17

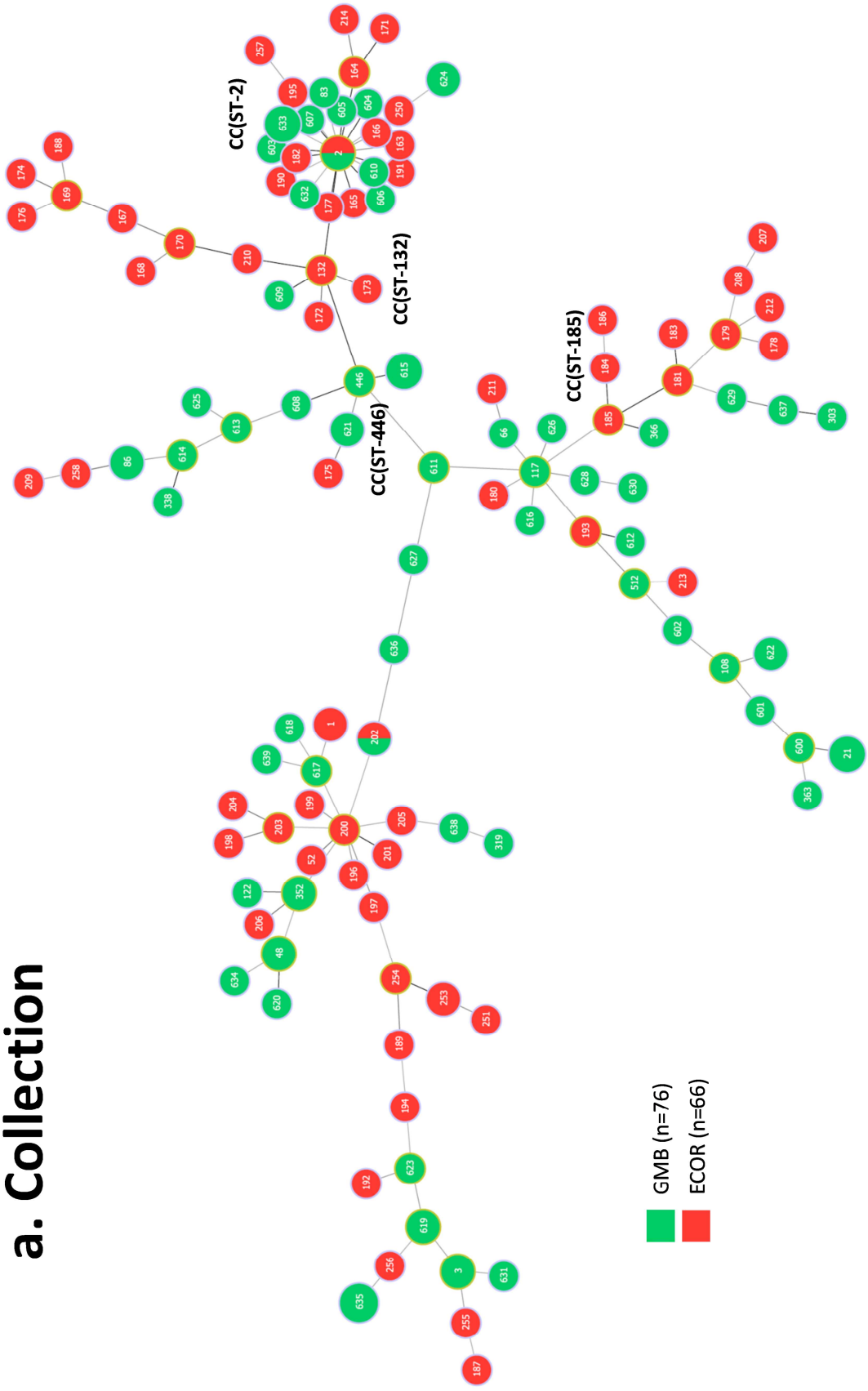

a. Collection
CC(ST-2)
CC(ST-132)
CC(ST-446)
CC(ST-185)
GMB (n=76)
ECOR (n=66)

Figure 3.17

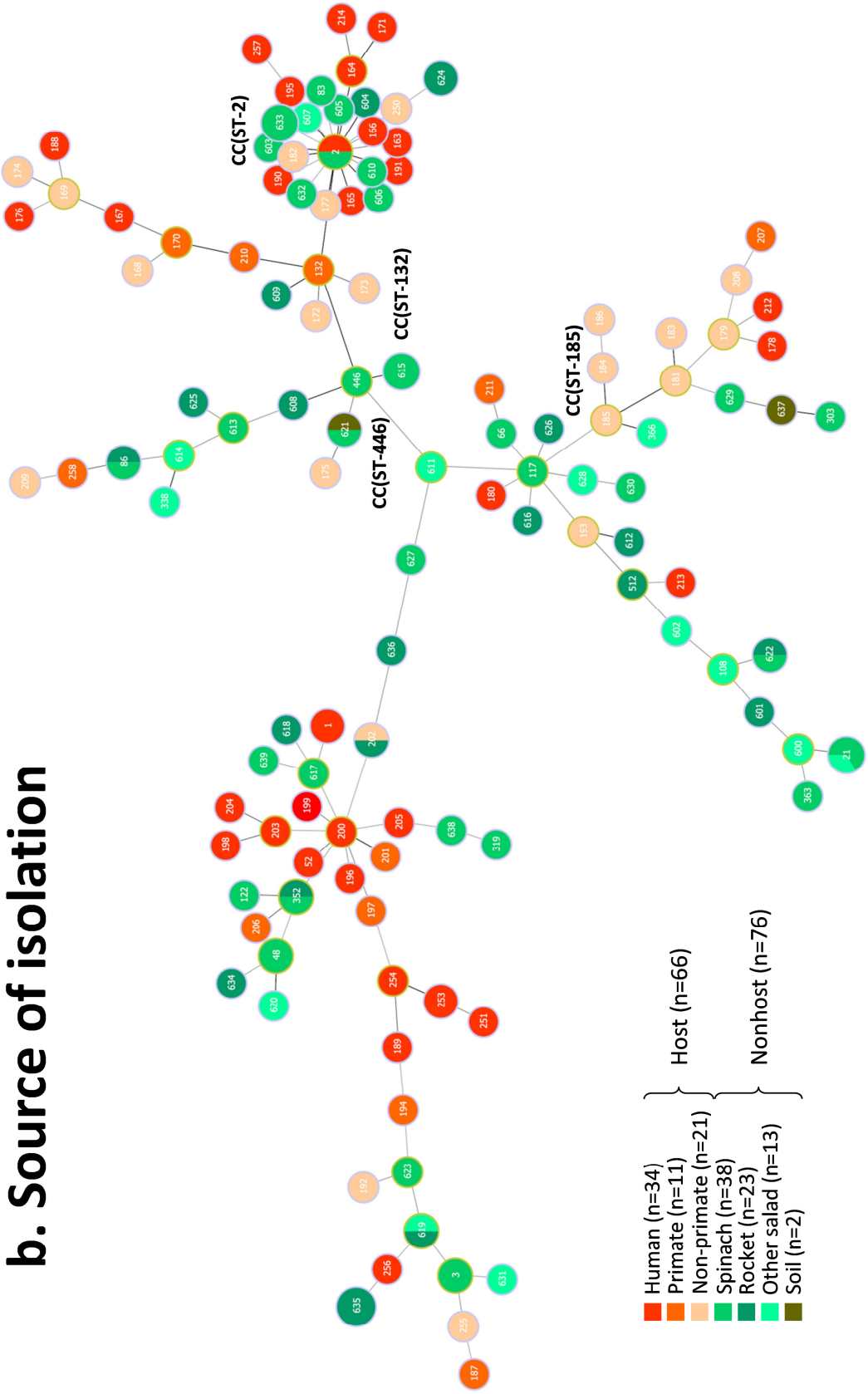

b. Source of isolation
CC(ST-2)
CC(ST-132)
CC(ST-446)
CC(ST-185)
Host (n=66)
Nonhost (n=76)
Human (n=34)
Primate (n=11)
Non-primate (n=21)
Spinach (n=38)
Rocket (n=23)
Other salad (n=13)
Soil (n=2)

Figure 3.17

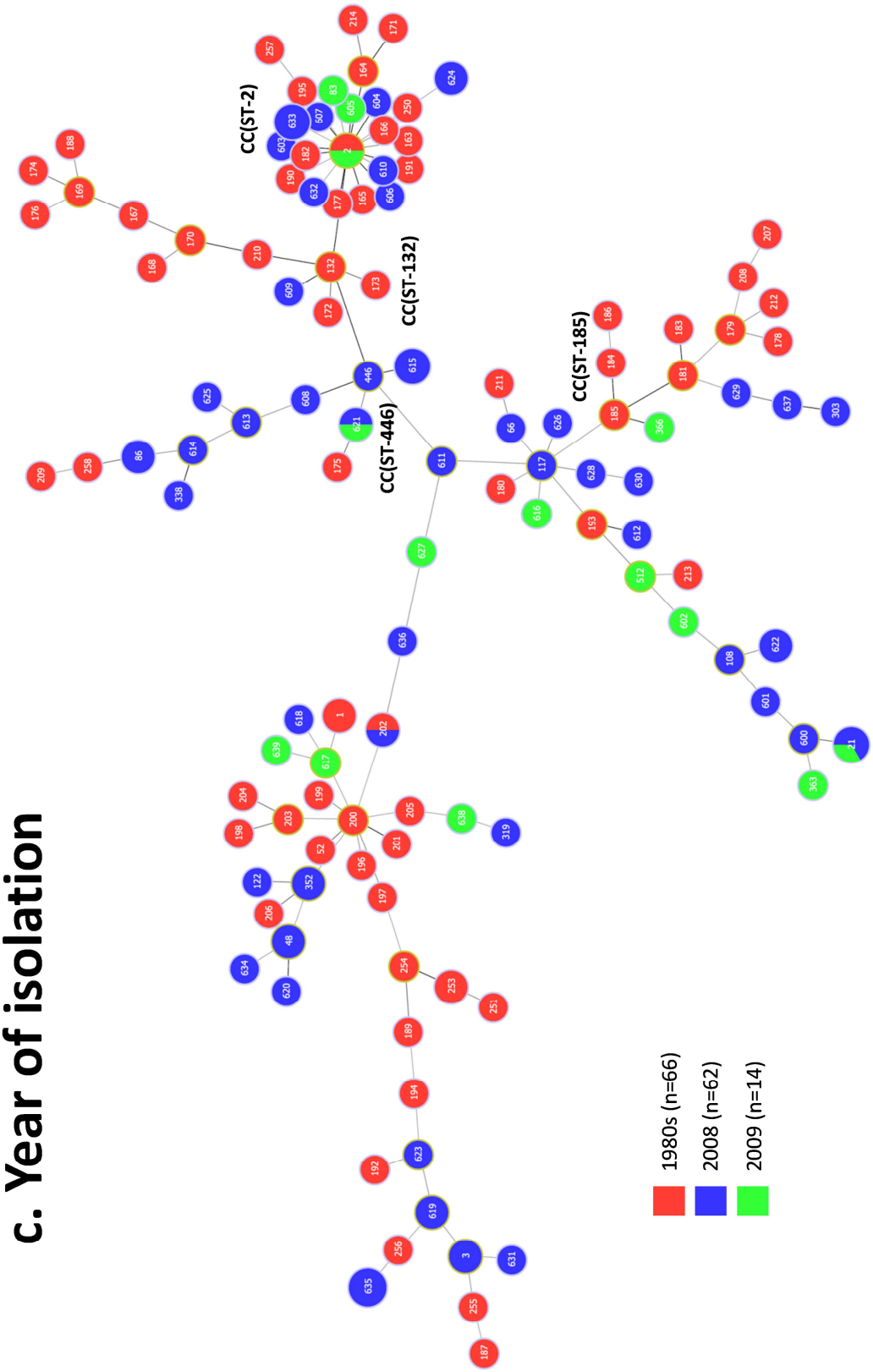

c. Year of isolation
CC(ST-2)
CC(ST-132)
CC(ST-446)
CC(ST-185)
1980s (n=66)
2008 (n=62)
2009 (n=14)

Figure 3.17

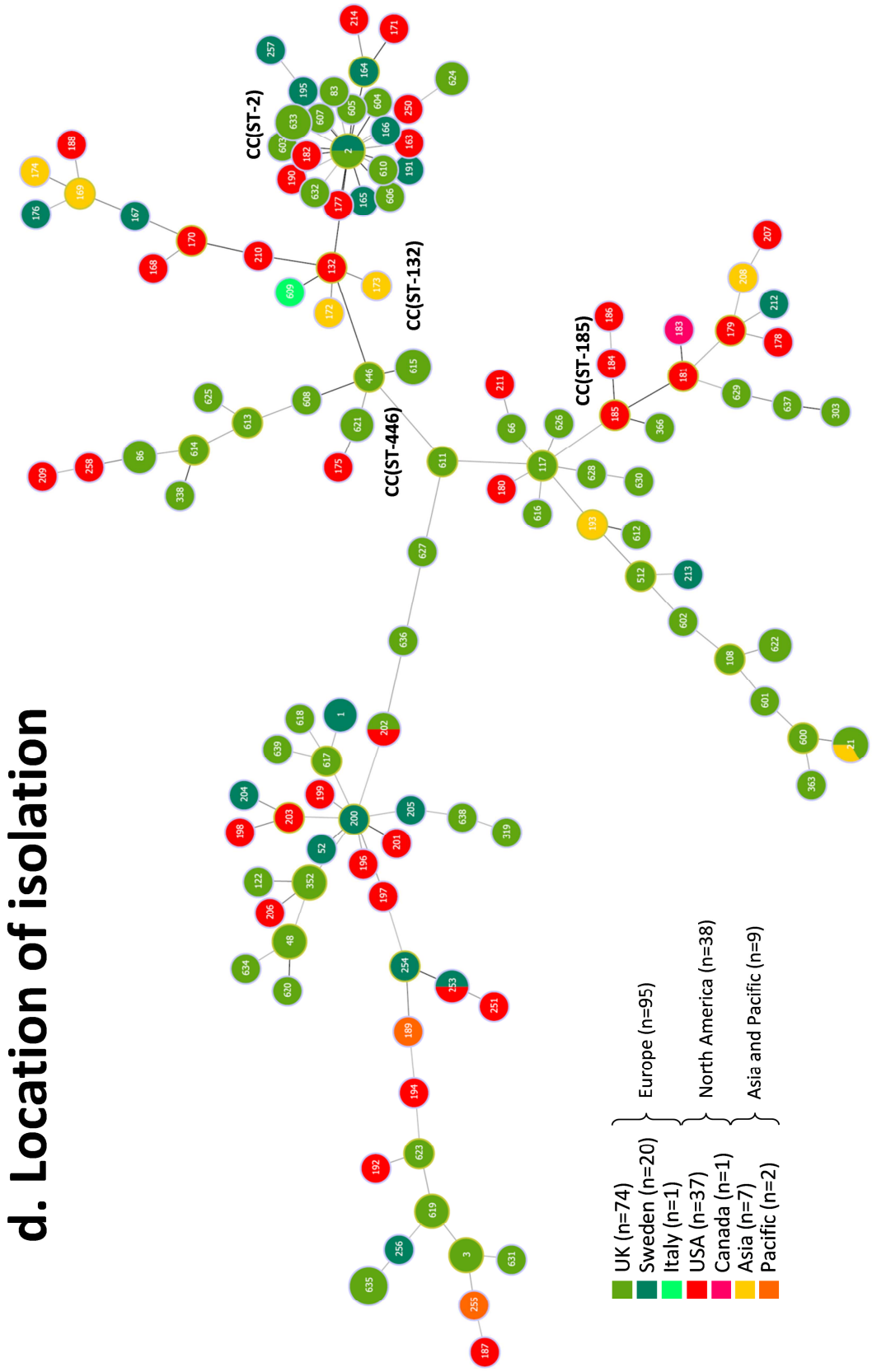
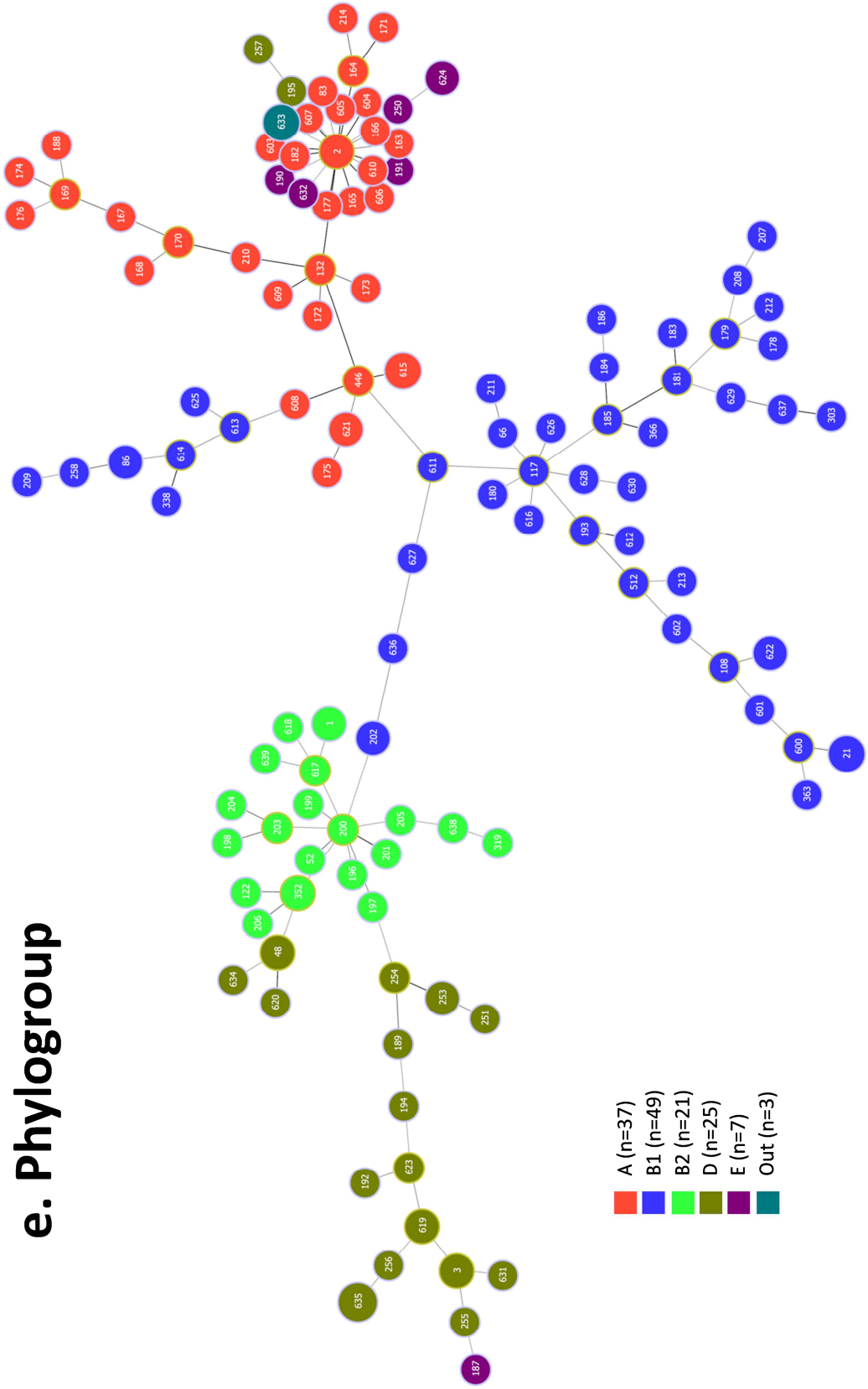

d. Location of isolation
CC(ST-2)
CC(ST-132)
CC(ST-446)
CC(ST-185)
UK (n=74)
Sweden (n=20)
Italy (n=1)
USA (n=37)
Canada (n=1)
Asia (n=7)
Pacific (n=2)
Europe (n=95)
North America (n=38)
Asia and Pacific (n=9)

Figure 3.17



According to the collection (GMB/ECOR) colouring (**Figure 3.17a**), we could generally observe ECOR STs that were related to GMB STs and vice versa. For instance, ST-2 was composed of both an ECOR and a GMB strain, and branching STs were found in both ECOR and GMB too. This suggests that host and nonhost isolates are not differentially positioned in the phylogeny of *E. coli*. There was also weak or absence of visible clustering according to source (**Figure 3.17b**), year (**Figure 3.17c**) and location (**Figure 3.17d**) of isolation, which is expected as these 3 properties are all directly linked to collection.

However, there were strong visible clusters according to phylogroups, confirming that our strains evolve in a quasi-clonal way. This result is expected, as phylogroup evolution is deeper than any other tested parameter for *E. coli* and has also been reported previously (Jaureguy et al., 2008). It suggests that the micro-evolution of STs is mostly constrained within phylogroups and confirms that *E. coli* is evolving in a semi-clonal way, forming rather distinct phylogenetic groups most of the time.

3.4.2.2. Construction of phylogenetic trees

BURST analyses and the construction of MS_{TREES} can only inform on the relatedness of sampled strains based on their allelic differences. The construction of phylogenetic trees adds the power of determining if two isolates share a common ancestor in the absence of this common ancestor among the samples (something impossible with minimum spanning analyses). The genealogy of strains can be reconstructed, and the resulting population structure is more precise than observed with MS_{TREES}.

There are multiple methods and algorithms available to reconstruct phylogenetic trees (Baldauf, 2003; Grunwald and Goss, 2011), but we chose to use ClonalFrame, a powerful Bayesian-based inference algorithm including an evolutionary model allowing the identification of clonal relationships between isolates by taking into account the recombination events that have disrupted the clonal inheritance. We produced a phylogenetic tree based on ClonalFrame output (see Material and methods for more details), using MEGA5 software (**Figure 3.18**). In the light of multiple and recent produce-related outbreaks, we also added pathogenic *E. coli* sequences to this analysis (**Figure 3.18**). We extracted the sequences required by the MLST scheme used in this study from the publicly available genomes of 4 *E. coli* strains that have been associated with produce-related outbreaks. Among these strains, *E. coli* O157:H7 strain Sakai was isolated in 1996 from a radish-related outbreak in Japan (Michino et al., 1999), *E. coli* O157:H7 strain TW14359 in 2006 from a spinach-related outbreak in USA (Kulasekara et al., 2009), and *E. coli* O104:H4 strains LB226692 and TY-2482 from a recent 2011 sprout-related outbreak in Germany (Rohde et al., 2011).

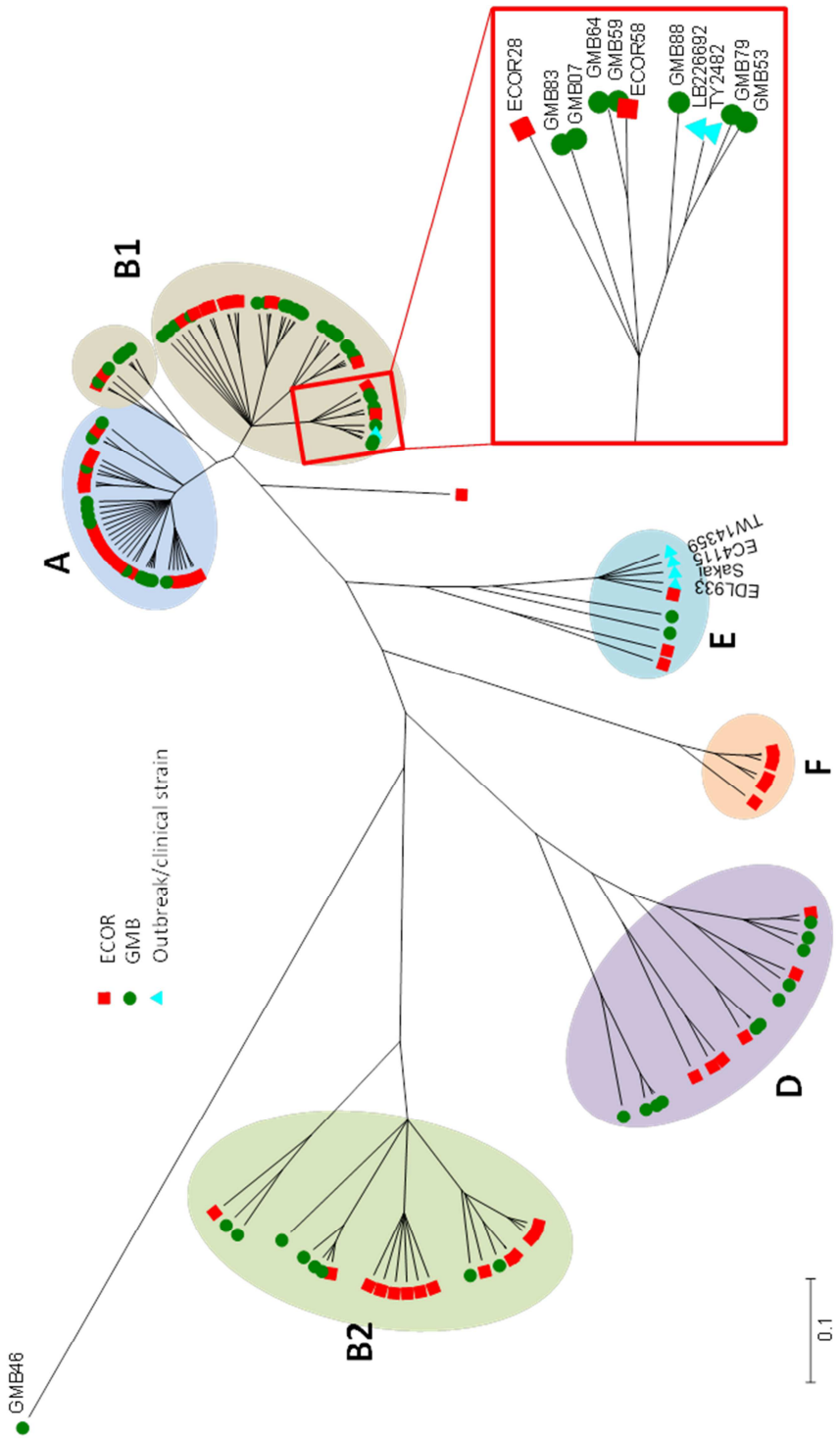


Figure 3.18.ClonalFrame phylogenetic tree showing evolutionary relationships between *E. coli* strains. The tree is based on MLST sequences of ECOR and GMB strains with the addition of 4 O157:H7 strains and 2 O104:H4 strains from public databases. Red squares indicate ECOR, green circles, GMB and blue triangles the pathogenic strains. The inset is a zoom on the phylogenetic neighbours of German outbreak O104:H4 strains.

The topology of the tree was consistent with the previously reported clonal-like structure of *E. coli* population. There were 4 major clades, or phylogenetic groups, called A, B1, B2 and D. There were GMB isolates in every major phylogenetic group, as observed with the triplex PCR assay. In every clade gathering GMB strains there were ECOR strains, confirming the diversity calculations. Minor clades regrouped phylogroup E and the recently described phylogroup F (Jaureguy et al., 2008), although no tested GMB strain clustered in it. Compared to phylogroups B2 and D, phylogroups A and B1 were closely related, suggesting a recent divergence and possibly higher levels of recombination.

There was an additional smaller clade diverging from phylogroup A and that we also considered as B1, as it is composed of ECOR isolates previously assigned to B1 using other schemes and of GMB strains assigned to B1 using the triplex PCR. This additional B1 clade has also been reported in earlier studies, and called the "ET-1 clade" in reference to its electrophoretic type in MLEE experiments (Walk et al., 2007). It has been found that isolates from the ET-1 clade were over-represented in natural environmental isolates, as observed with isolates from freshwater beaches (Walk et al., 2007). However, no phenotypical differences were observed for ET-1 isolates compared to the rest of B1 isolates, so there is no way to fully confirm that our observed additional B1 clade is indeed composed of ET-1 isolates.

As previously shown (Sims and Kim, 2011), isolates of serotype O157:H7 clustered within phylogroup E, along with 2 other GMB clones and ECOR37, which is the only ECOR isolate with a LEE locus and believed to be an O157:H7 progenitor from the O55:H7 serotype. O104:H4 strains from the recent produce-related outbreak in

Germany (Rohde et al., 2011) clustered in phylogroup B1, more specifically with ECOR28 (O104:H−), ECOR58 (O112:H8) and 7 other GMB clones. Apart from indicating that phylogenetic trees are inappropriate to make any clear assumption on the pathogenicity of isolates, this observation shows that strains from the same phylogenetic background as the recent German outbreak can be isolated on agricultural plants.

Three GMB clones (GMB46, 56, 57) were found to be distant from the rest, an observation already made in several *E. coli* population studies (Wirth et al., 2006; Walk et al., 2007; Walk et al., 2009; Luo et al., 2011). Earlier reports of strains identified biochemically as *E. coli* but phylogenetically distant have suggested their possible existence as ancestral variants of *E. coli,* remnants of an eventual selective sweep in *E. coli* that would have occurred 10 to 30 million years ago (Wirth, Falush et al. 2006). More recently, such phylogenetically distant isolates have been defined as members of cryptic lineages within the *Escherichia* genus, phylogenetically located between *E. coli* and its closest species *E. albertii* (Walk, Alm et al. 2009). To determine if our strains belonged to these cryptic lineages, we extracted sequences corresponding to our MLST scheme from the recently available genomes of representative strains from *Escherichia sp*. cryptic clades and *E. albertii* (Luo, Walk et al. 2011) and reconstructed a phylogenetic tree (**Figure 3.19**).

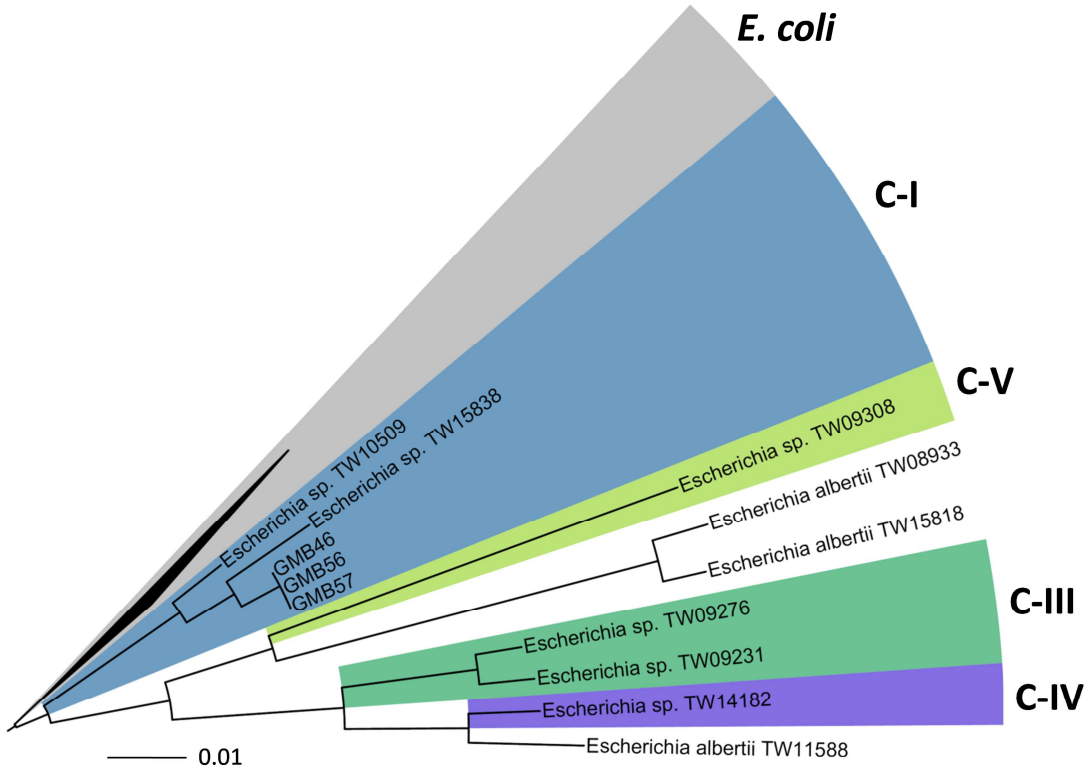


Figure 3.19. **Maximum likelihood phylogenetic tree showing relationships between *E. coli* GMB46, 56 and 57 with representatives strains from *Escherichia* cryptic lineages and *E. albertii***. The tree is based on MLST sequences from this work and from the recent sequences of *Escherichia sp*. and *E. albertii* strains (Luo, Walk et al. 2011). The tree was realised using iTOL (Letunic and Bork, 2011).

Strain representatives of the cryptic lineages within the *Escherichia sp*. genus have been recently described and sequenced (Walk, Alm et al. 2009, Luo, Walk et al. 2011, Clermont et al. 2011, Ingle et al. 2011). Prevalence has been examined, and it was hypothesised that these strains were environmentally-adapted and found in the majority in wild animals rather than humans (Clermont et al., 2011; Ingle et al., 2011). Our result is the first to report isolates of *Escherichia sp*. cryptic lineages on plants. Indeed, strains GMB46, GMB56 and GMB57 clustered with strains TW10509 and TW15838 (**Figure 3.19**), which are representatives from the Clade-I lineage, isolated from human faeces in India and freshwater sediments in Australia

respectively (Luo, Walk et al. 2011). These cryptic isolates have not been reported to strongly differ phenotypically from *E. coli* (Ingle, Clermont et al. 2011), and are therefore co-isolated with *E. coli* using standard isolation procedures, which led some to suggest that their presence could bias faecal indication tests (Walk, Alm et al. 2009). However, there is no convincing proof that the primary niche of these isolates is not the vertebrate gastrointestinal tract. On the contrary, their high degree of metabolic similarity with *E. coli* suggests it is and more studies are required to elucidate the ecology of these lineages.

The order of divergence of the different phylogroups has been a subject of debate since the first *E. coli* phylogenies were published, mainly in an effort to determine if the *E. coli* ancestor was pathogenic or not (Lecointre et al., 1998; Touchon et al., 2009; Sims and Kim, 2011). It seems probable that either B2 (Lecointre, Rachdi et al. 1998) or D (Touchon, Hoede et al. 2009) is the most ancestral group, whereas A and B1 diverged later and are considered as evolutionary "sister" clades. The use of unrooted radial visualisation of trees is not suitable to investigate the order of emergence of the different phylogroups (**Figure 3.19**). We thus produced a circular tree, rooted on GMB46, as it is evolutionarily more distant than the rest of the tested *E. coli* strains (**Figure 3.20**).

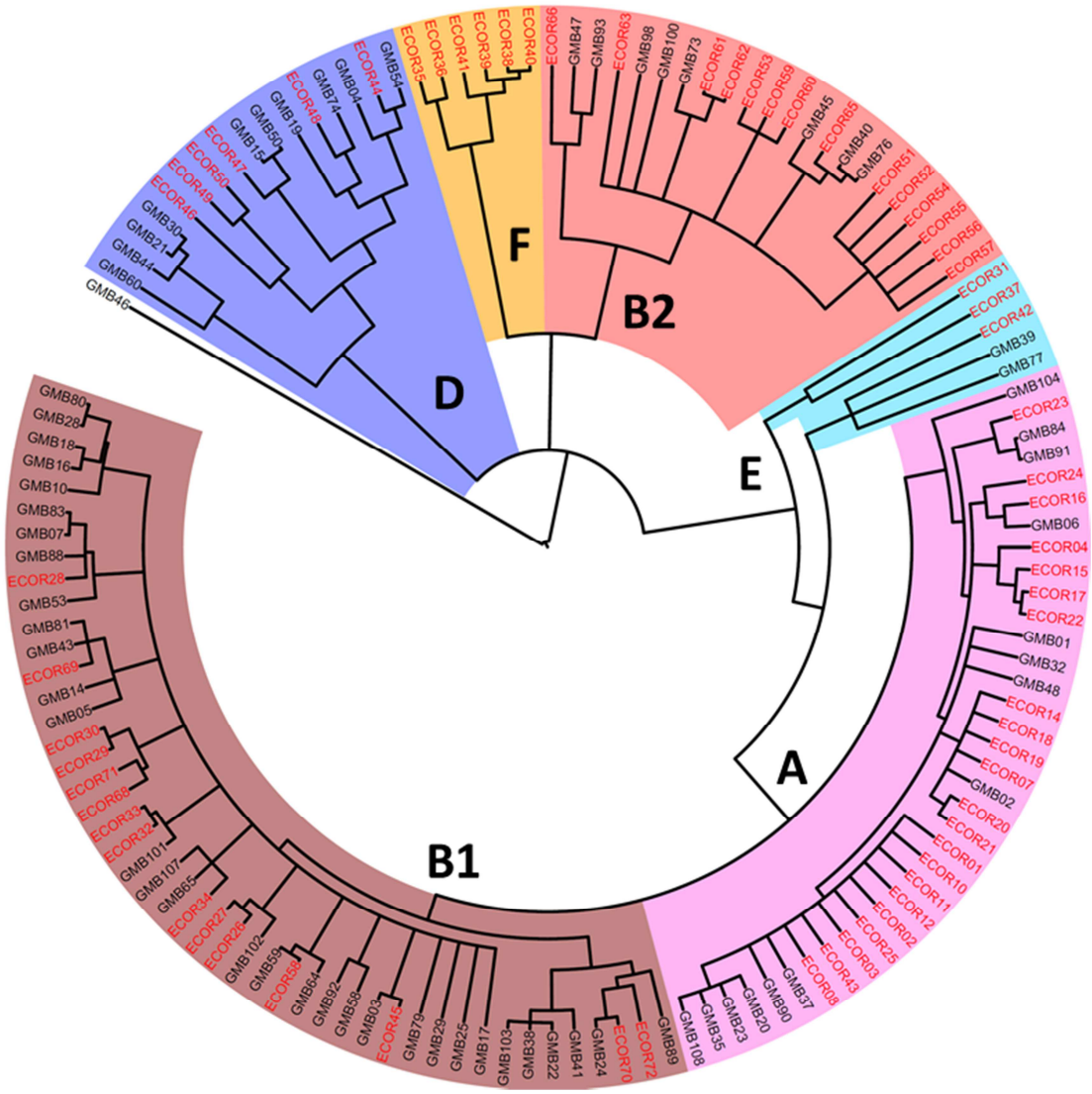


Figure 3.20. ClonalFrame phylogenetic tree regrouping plant-associated GMB isolates and host-associated ECOR strains. ECOR labels are written in red and GMB in black. The tree was visualised and annotated using iTOL (Letunic and Bork, 2011).

Using our MLST scheme, it is clear that phylogroup D appears at the bottom of the tree, and is likely to be the most ancestral phylogroup. No clade was paraphyletic except phylogroup E. A-B1 appeared as sister groups and to have diverged later. The newly described phylogroup F also seems to be related to phylogroup B2, which both appear to have diverged after D. Our observation of phylogroup emergence is globally

consistent with the findings reported (Touchon, Hoede et al. 2009, Sims and Kim 2011), and we find good support that phylogroup D rather than B2-F diverged first.

3.4.2.3. Examination of clonality structure of **E. coli** *populations based on our comparative analysis*

Before the discovery of the extent of homologous recombination and its role in shaping bacterial genomes and phenotypic properties within populations bacteria were assumed to evolve clonally, meaning that most of the observed variation among natural populations was assumed to be caused by mutation only. Recombination has been shown to be prevalent in most bacterial genomes, leading to a spectrum of population structures from highly clonal, such as monomorphic pathogens like *Salmonella enterica* serovar Typhi, *Mycobacterium tuberculosis* or *Yersinia pestis* (Achtman, 2008) to panmictic (or so called "epidemic") structures like *Neisseria sp.* (Maynard Smith et al., 1993) for which every clone is very different.

E. coli has been described to have a rather clonal population structure on the basis that distinct and stable phylogenetic groups were delimited by phylogenetic reconstruction methods based on multilocus analyses (Tenaillon et al., 2010). However, this hypothesis has been questioned by Wirth et al. (2006) who found that phylogenetic approaches were not statistically robust to accurately represent the population structure of organisms that often interbreed via homologous recombination, which seems to be the case for *E. coli* (Wirth, Falush et al. 2006). It was shown that *E. coli* housekeeping genes (in the context of MLST) were composed of an admixture (i.e., the interbreeding of different populations within a species) of multiple ancestral

groups, between which recombination seemed to have often occurred (Wirth et al. 2006). In fact, admixture analysis did not even clearly identify a phylogenetic background for isolates with highly admixed housekeeping genes (therefore classifying them as “ABD” or “AxB1” hybrids) even if those same isolates were clearly assigned to a phylogroup using traditional phylogenetic analyses (Wirth et al. 2006). Interestingly, this study also observed a tendency of *E. coli* pathogens to be among this hybrid population of admixed ancestry (Wirth et al. 2006). In the light of this discrepancy, some recent studies have preferred to classify isolates according to ancestry groups rather than phylogroups as defined traditionally by MLST and triplex PCR (Martinez-Medina et al., 2009). In this work, we kept using the classical phylogroup denominations (A, B1, B2, D, E and F) for comparison purposes with other studies.

Multiple factors can be examined to estimate the clonality of a population structure based on multilocus data. In a theoretically fully clonal population, there is no observable genetic exchange caused by recombination, and variation is solely caused by mutation. The first indication of a clonal population is to detect linkage disequilibrium in the tested population, which is the non-random association of alleles at different loci. Populations that are highly clonal exhibit a very high level of linkage disequilibrium, as their alleles at different loci are very similar (not randomly associated). On the other hand, populations of *Neisseria* are panmictic, implying that they show linkage equilibrium at multiple loci: every clone has a different allele (Maynard Smith et al., 1993; Didelot and Maiden, 2010). We detected significant linkage disequilibrium between the 8 housekeeping genes of our MLST experiment

on the whole *E. coli* dataset using calculations implemented in the START2 package (http://pubmlst.org/software/analysis/start2/).

A second property of clonal populations is a tree-like phylogeny, with distinct clades and a relatively high level of congruence between single gene phylogenies and the clonal genealogy of the population (Didelot and Maiden 2010). To investigate this, we created ClonalFrame single-gene phylogenies for each locus analysed using the Pasteur MLST scheme, and we calculated the congruence index (I_{cong}) between each of them and the 8-gene phylogeny (from **Figure 3.18**) using the method suggested by de Vienne et al. 2007 (**Table 3.6**). This test basically compares tree topologies between them and determines if the observed congruence between two tree topologies is higher than what would be expected by chance only (de Vienne et al., 2007).

Table 3.6. Topological congruence between single-gene phylogenies and the clonal genealogy of ECOR and GMB strains. MLST data was used and trees were constructed after a ClonalFrame analysis. Indices were calculated using the online tool created by de Vienne et al. 2007: http://www.ese.u-psud.fr/bases/upresa/pages/devienne/index.html.

Locus[a]	I_{cong}[b]	p-value	Congruence?
dinB	1.60330	3.38E-07	Yes
icdA	1.43151	2.66792E-05	Yes
pabB	0.85891	56.36239966	No
polB	1.43151	2.66792E-05	Yes
putP	1.08795	0.166365379	No
trpA	1.83234	9.97E-10	Yes
trpB	1.37425	0.00011446	Yes
uidA	0.85891	56.36239966	No

a. Gene analysed using ClonalFrame to produce a single-gene phylogenetic reconstruction, which was compared to the concatenated 8-genes phylogeny previously described.

b. Congruence index, as calculated using the method and online tool.

Single-gene phylogenies at 5 different loci showed phylogenetic congruence with the clonal genealogy of strains as determined by MLST (**Figure 3.18**) but the fact that 3 single-gene phylogenies (*pabB*, *putP* and *uidA*) were not congruent, suggests that *E. coli* (as summarised here by our ECOR and GMB strains) is not evolving in a highly clonal manner, providing that evolution at these 8 loci is a good representative of the population structure dynamics. This is consistent with the previous observations that *E. coli* clonality is not very strong (Wirth, Falush et al. 2006).

In the light of these observations, it becomes interesting to assess what is the impact of recombination on the population structure of our strains of *E. coli*. The model used for our ClonalFrame determination of phylogeny (**Figure 3.18**) represents the true clonal genealogy of strains independently of any noise on the phylogenetic signal. Conversely, there are ways to represent this recombinational noise on the phylogenetic signal, for instance by using phylogenetic networks for which no arbitrary decision is taken on the clustering of different leaves, as it can be the case with most phylogenetic reconstruction methods (neighbour-joining, maximum likelihood). In phylogenetic networks contrary to phylogenetic trees, leaves are linked by parallelograms rather than straight lines, indicating conflicting signals mainly due to recombination for which the algorithm was not able to decide a clear branch (Huson, 1998). In a way, phylogenetic networks show the opposite of ClonalFrame trees, that is to say very loose relationships between isolates, regardless of recombination. Based on concatenated sequences at 8 loci for our ECOR and GMB isolates, we used the SplitsTree4 software with the NeighbourNet algorithm to reconstruct a phylogenetic network (**Figure 3.21**).

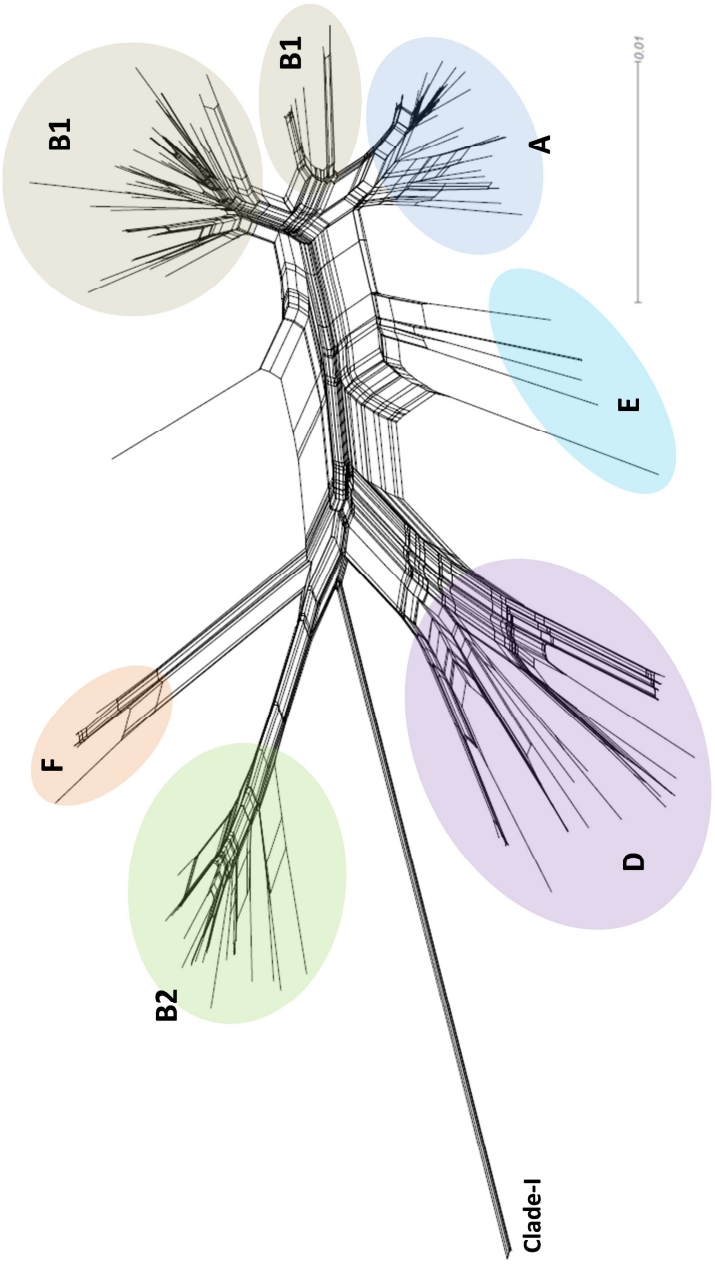


Figure 3.21. Phylogenetic network of ECOR and GMB strains based on concatenated MLST sequences at 8 loci. The network was constructed using SplitsTree4 and the NeighbourNet algorithm. Parallelograms denote incongruent phylogenies for particular branches most likely due to recombination.

There were high levels of recombination between phylogroups, as shown by the amount and size of parallelograms between each phylogroups instead of unambiguous linear branches (**Figure 3.21**). In spite of these high levels of recombination, clear phylogroups could nevertheless still be delimited, strengthening the fact that *E. coli* is evolving in a semi-clonal manner. Using this method, we observed that B2 had more highly congruent phylogenies than other phylogroups, as shown by the smaller parallelograms defining this clade (**Figure 3.21**). Clade-I isolates were still appearing to branch from B2, as were phylogroup F strains, something which was not observed on the ClonalFrame tree. There seemed to be similar levels of congruent phylogenies within phylogroup B1 and D, and we still observed an additional clade of B1 isolates. Indirectly, recombination then seemed slightly higher in phylogroup A, where more and bigger parallelograms were observed. Interestingly, this observation can be indirectly reflected in the ClonalFrame tree (**Figure 3.18**), where most of isolates in phylogroup A are placed at identical distances from their closest node. Indeed, because of high recombination levels in these strains, ClonalFrame cannot infer any satisfying clonal history between them, and thus places them at equidistance of the same node (**Figure 3.18**).

3.5. Comparative genomic hybridisation (CGH) to investigate genetic differences between collections or phylogroups

From the observations made in the previous section, it emerges that *E. coli* strains from various environments expectedly show clonal and genomic cohesion, and that signature traits of nonhost association in *E. coli* (if any) are not obvious when phylogenetic information is considered alone. In the last part of this chapter, we examined if there were any obvious differences in gene content between collections and phylogroups. We performed CGH on 21 GMB and 20 ECOR strains using ShEcoliO157 microarrays (see Methods). We used the Mann-Whitney-Wilcoxon (MWW) test with Bonferroni multiple testing correction to find gene content correlated with either collection or phylogroups.

We could not detect genes whose presence or absence was significantly correlated with the origin of isolation (GMB vs. ECOR). Removing the Bonferroni correction led to a list of 114 genes weakly associated with either GMB or ECOR strains. Among these weakly associated genes, the *phn* phosphonate metabolic cluster was notably more present in GMB than ECOR strains (data not shown), possibly suggesting that additional phosphorus metabolism is an important factor for life and persistence in secondary environments such as soils or plants. As the association is only weak using our conditions, the risk of false positives is high and we did not pursue this observation. No genes were found to be associated among GMB strains with the plant source of isolation, nor specifically in isolates from the same geographical location. Moreover, there were no specific genes associated with tested GMB strains isolated from the same field at the same time compared to the rest.

However, we identified 19 genes whose patterns of presence/absence were strongly associated with phylogenetic groups. As we had a higher proportion of strains from phylogroups A and B1 tested by CGH, association studies were based on the patterns of gene presence/absence in only 3 handmade groups: A, B1 and a third group composed of B2, D, E and F strains (**Table 3.7**).

Table 3.7. Genetic association with phylogroups based on ShEcoliO157 and CGH data.

ID	Name	Description	p-value	Distribution		
				A (n=11)	B1 (n=19)	B2-D-E-F (n=9)
b1268	*yciQ*	predicted inner membrane protein	0.00728	10 (90.9%)	0 (0%)	9 (100%)
Z5029	Z5029	putative adhesin	0.00728	1 (0.1%)	19 (100%)	0 (0%)
SF3640	SF3640	hypothetical protein	0.0211	2 (18.1%)	19 (100%)	0 (0%)
SF3641	SF3641	hypothetical protein	0.0276	1 (0.1%)	19 (100%)	1 (11.1%)
b0070	*yabM, setA*	broad specificity sugar efflux system	0.0479	11 (100%)	19 (100%)	0 (0%)
b0608	*ybdR*	predicted oxidoreductase, Zn-dependent and NAD(P)-binding	0.0479	11 (100%)	19 (100%)	0 (0%)
b0730	*farR, mngR*	mannosyl-D-glycerate transport/metabolism system repressor	0.0479	11 (100%)	19 (100%)	0 (0%)
b0731	*hrsA, mngA*	fused 2-O-A-mannosyl-D-glycerate specific PTS enzymes: IIA component/IIB component/IIC component	0.0479	11 (100%)	19 (100%)	0 (0%)
b2339	*yfcV*	predicted fimbrial-like adhesin protein	0.0479	11 (100%)	19 (100%)	0 (0%)
b3143	*yraI*	predicted periplasmic pilin chaperone	0.0479	11 (100%)	19 (100%)	0 (0%)
b3145	*yraK*	predicted fimbrial-like adhesin protein	0.0479	11 (100%)	19 (100%)	0 (0%)
VIREC O255	*chuA*	outer membrane heme/hemoglobin receptor	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4910	*chuS*	putative heme/hemoglobin transport protein	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4911	*chuA*	outer membrane heme/hemoglobin receptor	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4913	*chuT*	iron complex transport system substrate-binding protein	0.0479	0 (0%)	0 (0%)	9 (100%)

Z4914	*chuW*	oxygen-independent coproporphyrinogen III oxidase	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4915	*chuX*	hypothetical protein	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4917	*chuY*	hypothetical protein	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4918	*chuU*	putative permease of iron compound ABC transport system	0.0479	0 (0%)	0 (0%)	9 (100%)
Z4919	*hmuV*	hemin importer ATP-binding subunit	0.0479	0 (0%)	0 (0%)	9 (100%)

One gene, *yciQ*, encoding a putative inner membrane protein, was specifically absent from B1 strains only, as already previously reported (Touchon et al. 2009); whereas 3 genes including Z5029 encoding a putative adhesin were specifically present in B1 strains only (**Table 3.7**). However, Z5029 has previously been reported to be a pseudogene following analyses of two previously sequenced *E. coli* strains (Touchon et al. 2009), indicating a possible ongoing process of gene loss within the B1 phylogroup. Seven genes were present only in A and B1 strains, 4 of which are involved in metabolism and 3 are predicted adhesion factors (**Table 3.7**). Conversely, the whole *chu-hmuV* locus involved in heme metabolism was absent from all tested A and B1 strains (**Table 3.7**) including *chuA*, a gene used as a marker to identify strains from phylogroups B2 and D by the triplex PCR method (Clermont et al. 2000).

3.6. Industrial relevance

It is a question of growing interest for food industry and farmers to know if all *E. coli* isolates found on plants during routine tests are similar or not. Our results provide the first report that isolates representing the 4 major phylogenetic groups A, B1, B2 and D can be found in abundance in agricultural settings and on plants. This observation suggests that field contamination is a non-uniform event, likely originating from multiple and complex contamination sources, in line with the hypothesis that agricultural fields are contaminated mainly by irrigation or manure application and not so much by direct faecal contamination (See Introduction section 1.3). Moreover, our observation that isolates from one single field on the same day are statistically as diverse as isolates from the whole UK further support this observation. As *E. coli* strains able to be transferred via water or soil are likely to be of diverse origins themselves, it is not surprising to observe high phylogenetic diversity in the resulting retrieved plant-associated population. This also indicates that the selection pressure on *E. coli* by environments linked to agricultural settings (water, soil or plants) is not strong enough to drive the selection for specific strains, implying that very diverse *E. coli* isolates can potentially colonize agricultural soils and crops. Nonetheless, among plant-associated isolates we found B1 isolates in proportions far superior to those of phylogroup A, B2 or D, suggesting that within the observed phylogenetic diversity of *E. coli* from plants, a subset of strains could possess a higher fitness on plants than the rest, and thus be retrieved at higher levels.

Previous literature on population structures of nonhost isolates of *E. coli* indicated that strains from phylogroup B1 were also isolated in majority from soils (Bergholz et al., 2011) and freshwaters (Walk, Alm et al. 2007). Interestingly, strains from B1 were also isolated in greater numbers from fish, frogs, reptiles, birds and carnivorous mammals in an Australian survey (Gordon et al. 2008) (**Table 1.1**), from healthy and diseased cattle in France, United States, Brazil and Iran (**Table 1.1**) and pigs in France (Bibbal et al., 2009). It is therefore plausible that a higher incidence in wildlife and farm animal faecal depositions coupled with increased nonhost persistence could be the cause of phylogroup B1 isolates being retrieved at higher levels from environmental nonhost sources. Based on our observations and the available literature, we formulate here the hypothesis that B1 is a nonhost-associated phylogroup. It is thus found mainly in environments outside *E. coli* either because it tends to harbour more traits enhancing nonhost fitness or simply because it is more prevalent in non-human hosts, especially in hosts of agricultural concern. In any case, faecal contamination on plants can be considered as non-direct and it is likely that plant isolates have undergone multiple nonhost-associated selection processes before colonizing plants.

An interesting application results from the observation of this specific population structure associated with plants. When comparing the proportions of phylogroups from different secondary environments (from other published studies) in a radar plot (**Figure 3.22**), we can observe a slight variation between the distributions in soil and water. When our newly described *E. coli* phylogroup distribution from plants is superimposed, it is almost identical to the distribution in water (**Figure 3.22**),

suggesting that the ecological constraints on *E. coli* life on water produced the same population structure as the one observed on plants.

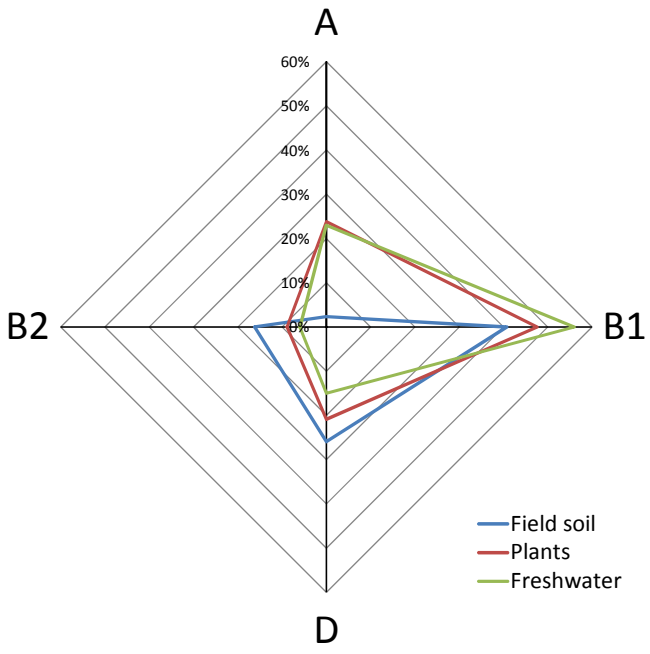


Figure 3.22. Radar plot of *E. coli* phylogroup distribution in three different secondary environments. The soil and freshwater data were taken from published studies (Bergholz, Noar et al. 2011; Walk, Alm et al. 2007) and the plant data from this study.

To simply conclude from this meta-analysis that irrigation contributed more significantly to *E. coli* contamination on plants is a big leap and requires confirmation by controlled field-scale experiments (although selective pressures on freshwater should theoretically be similar everywhere on Earth, the water study was conducted in Michigan, USA and our study in UK). However, this novel type of approach could prove useful, as previous studies on microbial source tracking (or MST) were always conducted on single strains, and their relatedness with strains from a known source. By looking at the population structure, rather than single strains to identify sources of contamination, the bias introduced by genomic rearrangements and the intrinsic genomic variability within *E. coli* is greatly reduced. This approach was tested to identify the source of sewage contamination by *E. coli* using population structures from various farm animals, with very promising results (Carlos et al. 2010).

4. Phenotypic variability between plant and faecal isolates of *E. coli* as a reflection of host and nonhost association

4.1. Context

In the previous section, we characterised the phylogenetic relationships and the population structure of plant-associated strains of *E. coli* and compared them to the faecal isolates of the ECOR collection. Apart from an unbalanced distribution of phylogroups showing a majority of plant strains from phylogroup B1, we could not find any clear differences at the phylogenetic level that would distinguish plant isolates from host-associated *E. coli*. This observation is consistent with earlier work supposing that the presence of *E. coli* in nonhost environment is caused by constant faecal fluxes, balanced by relatively rapid death (Winfield and Groisman, 2003). Nevertheless, *E. coli* strains do not seem equally able to survive and persist in the environment, as illustrated by the wide range of observations in survival studies using different strains (Whipps et al., 2008). It is possible that this variability is reflected in our sampling of plant isolates (GMB strains), which presumably have had a variable life history from faecal excretion before being retrieved on plants and thus have potentially resisted various earlier selection pressures, hopefully enriching for observable traits.

In this section, we focused on characterising the possible phenotypical differences existing between plant and host-associated isolates of *E. coli.* To narrow the range of phenotypes to assess, as suggested in the Introduction section 1.3.2, we hypothesised that metabolic abilities and phenotypes associated with colonisation in nonhost

environments to be key factors in environmental persistence and survival. We therefore used metabolic profiling using Biolog GN2 plates on GMB and ECOR strains and characterised their ability to form biofilms, produce siderophores and swim *in vitro*.

4.2. Variability in carbon metabolic profiling of plant and host strains of *E. coli*

4.2.1. Utilisation of Biolog: principle, controls and threshold determination

4.2.1.1. Principle of Biolog

The screening for metabolic abilities determines if a given strain has the potential to use nutrients for respiration and/or growth. In this section, we used 96-well plates manufactured by Biolog (Techno-Path, UK) to simultaneously assess growth on 95 different C-sources from the same inoculum. The plates used are of the GN2 type, initially designed to allow identification of Gram-negative bacteria. The substrate panel of GN2 plates has been designed to provide maximum discrimination between Gram-negative bacteria, is well utilised by typical *E. coli* strains. Six different types of substrates are represented on the GN2 type of plates we used (**Table 4.1**).

Table 4.1.C-sources on Biolog GN2 plates.

Well	Name	Chemical formula	Substrate guild[a]
A01	Water	H_2O	Control
A02	α-Cyclodextrin	$C_{36}H_{60}O_{30}$	Polymers
A03	Dextrin	$C_6H_{10}O_5$	Polymers
A04	Glycogen	$(C_6H_{10}O_5)_n$	Polymers
A05	Tween 40	$C_{26}H_{50}O_3$	Polymers
A06	Tween 80	$C_{64}H_{124}O_{26}$	Polymers
A07	N-Acetyl-D-Galactosamine	$C_8H_{15}NO_6$	Carbohydrates
A08	N-Acetyl-D-Glucosamine	$C_8H_{15}NO_6$	Carbohydrates
A09	Adonitol	$C_5H_{12}O_5$	Carbohydrates
A10	L-Arabinose	$C_5H_{10}O_5$	Carbohydrates
A11	D-Arabitol	$C_5H_{12}O_5$	Carbohydrates
A12	D-Cellobiose	$C_{12}H_{22}O_{11}$	Carbohydrates
B01	*i*-Erythritol	$C_4H_{10}O_4$	Carbohydrates
B02	D-Fructose	$C_6H_{12}O_6$	Carbohydrates
B03	L-Fucose	$C_6H_{12}O_5$	Carbohydrates
B04	D-Galactose	$C_6H_{12}O_6$	Carbohydrates
B05	Gentiobiose	$C_{12}H_{22}O_{11}$	Carbohydrates
B06	α-D-Glucose	$C_6H_{12}O_6$	Carbohydrates
B07	*m*-Inositol	$C_6H_{12}O_6$	Carbohydrates
B08	α-D-Lactose	$C_{12}H_{22}O_{11}$	Carbohydrates
B09	Lactulose	$C_{12}H_{22}O_{11}$	Carbohydrates
B10	Maltose	$C_{12}H_{22}O_{11}$	Carbohydrates
B11	D-Mannitol	$C_6H_{14}O_6$	Carbohydrates
B12	D-Mannose	$C_6H_{12}O_6$	Carbohydrates
C01	D-Melibiose	$C_{12}H_{22}O_{11}$	Carbohydrates
C02	β-Methyl-D-Glucoside	$C_7H_{14}O_6$	Carbohydrates
C03	D-Psicose	$C_6H_{12}O_6$	Carbohydrates
C04	D-Raffinose	$C_{18}H_{32}O_{16}$	Carbohydrates
C05	L-Rhamnose	$C_6H_{12}O_5$	Carbohydrates
C06	D-Sorbitol	$C_6H_{14}O_6$	Carbohydrates
C07	Sucrose	$C_{12}H_{22}O_{11}$	Carbohydrates
C08	D-Trehalose	$C_{12}H_{22}O_{11}$	Carbohydrates
C09	Turanose	$C_{12}H_{22}O_{11}$	Carbohydrates
C10	Xylitol	$C_5H_{12}O_5$	Carbohydrates
C11	Pyruvic Acid Methyl Ester	$C_4H_6O_3$	Miscellaneous
C12	Succinic Acid Mono-Methyl-Ester	$C_5H_8O_4$	Miscellaneous
D01	Acetic Acid	$C_2H_4O_2$	Carboxylic acids
D02	*Cis*-Aconitic Acid	$C_6H_6O_6$	Carboxylic acids
D03	Citric Acid	$C_6H_8O_7$	Carboxylic acids
D04	Formic Acid	CH_2O_2	Carboxylic acids

D05	D-Galactonic Acid γ-Lactone	$C_6H_{10}O_6$	Carboxylic acids
D06	D-Galacturonic Acid	$C_6H_{10}O_7$	Carboxylic acids
D07	D-Gluconic Acid	$C_6H_{12}O_7$	Carboxylic acids
D08	D-Glucosaminic Acid	$C_6H_{13}NO_6$	Carboxylic acids
D09	D-Glucuronic Acid	$C_6H_{10}O_7$	Carboxylic acids
D10	α-Hydroxybutyric Acid	$C_4H_8O_3$	Carboxylic acids
D11	β-Hydroxybutyric Acid	$C_4H_8O_3$	Carboxylic acids
D12	γ-Hydroxybutyric Acid	$C_4H_8O_3$	Carboxylic acids
E01	*p*-HydroxyPhenylacetic Acid	$C_8H_8O_3$	Carboxylic acids
E02	Itaconic Acid	$C_5H_6O_4$	Carboxylic acids
E03	α-Keto Butyric Acid	$C_4H_6O_3$	Carboxylic acids
E04	α-KetoGlutaric Acid	$C_5H_8O_5$	Carboxylic acids
E05	α-KetoValeric Acid	$C_5H_8O_3$	Carboxylic acids
E06	D,L-Lactic Acid	$C_3H_6O_3$	Carboxylic acids
E07	Malonic Acid	$C_3H_4O_4$	Carboxylic acids
E08	Propionic Acid	$C_6H_6O_2$	Carboxylic acids
E09	Quinic Acid	$C_7H_{12}O_6$	Carboxylic acids
E10	D-Saccharic Acid	$C_6H_{10}O_8$	Carboxylic acids
E11	Sebacic Acid	$C_{10}H_{18}O_4$	Carboxylic acids
E12	Succinic Acid	$C_4H_6O_4$	Carboxylic acids
F01	Bromosuccinic Acid	$C_4H_5O_4Br$	Miscellaneous
F02	Succinamic Acid	$C_4H_7NO_3$	Amines/amides
F03	Glucuronamide	$C_6H_{11}NO_6$	Amines/amides
F04	L-Alaninamide	$C_3H_8N_2O$	Amines/amides
F05	D-Alanine	$C_3H_7NO_2$	Amino acids
F06	L-Alanine	$C_3H_7NO_2$	Amino acids
F07	L-Alanylglycine	$C_5H_{10}N_2O_3$	Amino acids
F08	L-Asparagine	$C_4H_8N_2O_3$	Amino acids
F09	L-Aspartic Acid	$C_4H_7NO_4$	Amino acids
F10	L-Glutamic Acid	$C_5H_9NO_4$	Amino acids
F11	Glycyl-L-aspartic Acid	$C_6H_{10}N_2O_5$	Amino acids
F12	Glycyl-L-Glutamic Acid	$C_7H_{12}N_2O_5$	Amino acids
G01	L-Histidine	$C_6H_9N_3O_2$	Amino acids
G02	Hydroxy-L-Proline	$C_5H_9NO_3$	Amino acids
G03	L-Leucine	$C_6H_{13}NO_2$	Amino acids
G04	L-Ornithine	$C_5H_{12}N_2O_2$	Amino acids
G05	L-Phenylalanine	$C_9H_{11}NO_2$	Amino acids
G06	L-Proline	$C_5H_9NO_2$	Amino acids
G07	L-Pyroglutamic Acid	$C_5H_7NO_3$	Amino acids
G08	D-Serine	$C_3H_7NO_3$	Amino acids
G09	L-Serine	$C_3H_7NO_3$	Amino acids
G10	L-Threonine	$C_4H_9NO_3$	Amino acids
G11	D,L-Carnitine	$C_7H_{15}NO_3$	Amino acids
G12	γ-Amino Butyric Acid	$C_4H_9NO_2$	Amino acids

H01	Urocanic Acid	$C_6H_6N_2O_2$	Miscellaneous
H02	Inosine	$C_{10}H_{12}N_4O_5$	Miscellaneous
H03	Uridine	$C_9H_{12}N_2O_6$	Miscellaneous
H04	Thymidine	$C_{10}H_{14}N_2O_5$	Miscellaneous
H05	Phenyethylamine	$C_8H_{11}N$	Amines/amides
H06	Putrescine	$C_4H_{12}N_2$	Amines/amides
H07	2-Aminoethanol	C_2O_7NO	Amines/amides
H08	2,3-Butanediol	$C_4H_{10}O_2$	Miscellaneous
H09	Glycerol	$C_3H_8O_3$	Miscellaneous
H10	D,L-α-Glycerol Phosphate	$C_3H_9O_6P$	Miscellaneous
H11	α-D-Glucose-1-Phosphate	$C_6H_{13}O_9P$	Miscellaneous
H12	D-Glucose-6-Phosphate	$C_6H_{13}O_9P$	Miscellaneous

a. Substrate guilds were defined according to previously published groups (Preston-Mafham et al., 2002)

It has been estimated that several hundred mutations in *E. coli* genomes could alter the profiles obtained using Biolog 96-well plates (Cooper and Lenski, 2000) suggesting that Biolog plates are a suitable way to examine *E. coli* intra-species diversity using large collections of isolates. The Biolog system is based on simple redox chemistry (Bochner 2009). In each well, the C-source and bacterial inoculum are provided with a rich proprietary medium containing additional nutrients required for growth but excluding any additional C-source. This medium also contains tetrazolium chloride (or TTC), a redox indicator. When the C-source is used as an aerobic substrate by cells, NADH is produced during respiration, creating a reducing power in the well which leads to the irreversible reduction of tetrazolium to a colourful purple dye (Bochner 2009). Two phenotypes can be assessed using Biolog plates: respiration (the ability to use aerobically the corresponding C-source) and growth (the ability to replicate using the corresponding C-source). The two are almost always linked, although respiration can usually be detected before growth (Bochner 2009). Biolog plates are either available as 20 full panels of more than 2000 biochemical tests or, as we chose for costing reasons, as individual 96-well plates for

bacterial identification or small-scale comparison purposes. Biolog also commercialises the Omnilog system, a multiplex incubator with individual CCD cameras for each plate, monitoring the kinetics of tetrazolium reduction over time. Again, for costing reasons, we used a classical spectrophotometer to measure optical density (OD_{600}) in the plates, and not respiration as indicated by tetrazolium reduction.

4.2.1.2. Experimental controls

To control that we correctly followed the procedures, we compared the GN2 profiles we obtained with 2 reference strains (K-12 strain MG1655 and O157:H7 strain Sakai) with previously published literature (Mukherjee et al., 2008) for 4 different C-sources (**Figure 4.1**).

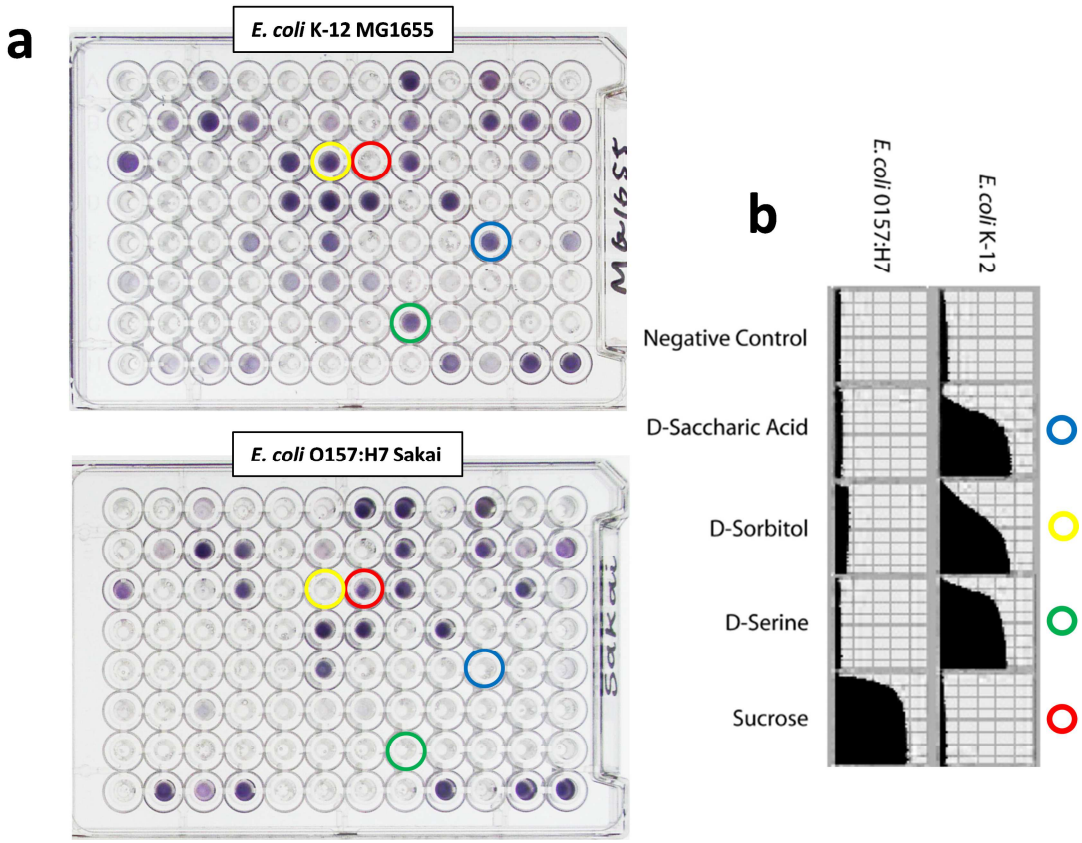


Figure 4.1.Visual controls of Biolog GN2 procedure. (a) Pictures of *E. coli* K-12 strain MG1655 (top) and *E. coli* O157:H7 strain Sakai (bottom) Biolog GN2 plates after 24 hours of inoculation at 37°C; (b) Previously published Omnilog kinetics data for the same strains (Mukherjee, Mammel et al. 2008). Coloured circles indicate correspondence between results.

We observed the same result as in published literature: positive utilisation of D-saccharic acid, D-sorbitol and D-serine for *E. coli* K-12 and not for O157:H7, and positive utilisation of sucrose for O157:H7 and not for K-12 (**Figure 4.1**). We assessed this assay reproducibility performing two independent biological replicates for 13 randomly chosen strains in our collections (**Figure 4.2**).

	ECOR10-a	ECOR18-a	ECOR32-a	ECOR38-a	ECOR57-a	ECOR67-a	ECOR72-a	GMB07-a	GMB23-a	GMB52-a	GMB76-a	GMB83-a	H1-1218-a	ECOR10-b	ECOR18-b	ECOR32-b	ECOR38-b	ECOR57-b	ECOR67-b	ECOR72-b	GMB07-b	GMB23-b	GMB52-b	GMB76-b	GMB83-b	H1-1218-b
ECOR10-a		0.72	0.76	0.74	0.83	0.84	0.79	0.70	0.73	0.71	0.80	0.78	0.74	0.94	0.69	0.76	0.75	0.81	0.83	0.76	0.69	0.72	0.64	0.79	0.73	0.71
ECOR18-a	0.72		0.66	0.58	0.74	0.74	0.65	0.65	0.72	0.69	0.70	0.68	0.64	0.70	0.79	0.63	0.62	0.68	0.76	0.66	0.66	0.73	0.62	0.72	0.75	0.71
ECOR32-a	0.76	0.66		0.85	0.82	0.88	0.85	0.84	0.65	0.74	0.80	0.89	0.70	0.72	0.66	0.90	0.85	0.80	0.83	0.84	0.84	0.65	0.68	0.85	0.83	0.81
ECOR38-a	0.74	0.58	0.85		0.75	0.82	0.79	0.74	0.72	0.77	0.77	0.79	0.64	0.71	0.60	0.84	0.96	0.79	0.78	0.73	0.73	0.72	0.73	0.81	0.73	0.75
ECOR57-a	0.83	0.74	0.82	0.75		0.87	0.77	0.69	0.74	0.72	0.78	0.79	0.73	0.81	0.69	0.76	0.74	0.83	0.85	0.73	0.68	0.73	0.68	0.81	0.72	0.69
ECOR67-a	0.84	0.74	0.88	0.82	0.87		0.84	0.83	0.76	0.77	0.86	0.89	0.76	0.81	0.73	0.84	0.82	0.84	0.97	0.81	0.83	0.76	0.70	0.90	0.85	0.79
ECOR72-a	0.79	0.65	0.85	0.79	0.77	0.84		0.84	0.62	0.75	0.81	0.85	0.73	0.76	0.66	0.83	0.76	0.74	0.79	0.95	0.83	0.61	0.68	0.81	0.82	0.78
GMB07-a	0.70	0.65	0.84	0.74	0.69	0.83	0.84		0.62	0.70	0.74	0.89	0.72	0.67	0.66	0.84	0.75	0.79	0.78	0.83	0.98	0.63	0.66	0.81	0.88	0.83
GMB23-a	0.73	0.72	0.65	0.72	0.74	0.76	0.62	0.62		0.69	0.67	0.69	0.71	0.70	0.78	0.69	0.74	0.71	0.74	0.59	0.62	0.99	0.60	0.69	0.67	0.64
GMB52-a	0.71	0.69	0.74	0.77	0.72	0.77	0.75	0.70	0.69		0.77	0.72	0.66	0.69	0.67	0.76	0.75	0.72	0.75	0.70	0.69	0.69	0.95	0.82	0.74	0.76
GMB76-a	0.80	0.70	0.80	0.77	0.78	0.86	0.81	0.74	0.67	0.77		0.79	0.68	0.78	0.70	0.76	0.78	0.75	0.80	0.78	0.75	0.67	0.69	0.93	0.78	0.74
GMB83-a	0.78	0.68	0.89	0.79	0.79	0.89	0.85	0.89	0.69	0.72	0.79		0.75	0.75	0.72	0.86	0.81	0.82	0.84	0.84	0.87	0.69	0.66	0.83	0.88	0.82
H1-1218-a	0.74	0.64	0.70	0.64	0.73	0.76	0.73	0.72	0.71	0.66	0.68	0.75		0.71	0.64	0.69	0.64	0.70	0.74	0.73	0.72	0.72	0.60	0.70	0.77	0.74
ECOR10-b	0.94	0.70	0.72	0.71	0.81	0.81	0.76	0.67	0.70	0.69	0.78	0.75	0.71		0.68	0.72	0.72	0.77	0.80	0.72	0.66	0.70	0.62	0.77	0.72	0.66
ECOR18-b	0.69	0.79	0.66	0.60	0.69	0.73	0.66	0.66	0.78	0.67	0.70	0.72	0.64	0.68		0.65	0.64	0.65	0.71	0.67	0.66	0.77	0.62	0.67	0.71	0.67
ECOR32-b	0.76	0.63	0.90	0.84	0.76	0.84	0.83	0.84	0.69	0.76	0.76	0.86	0.69	0.72	0.65		0.82	0.76	0.80	0.79	0.82	0.69	0.71	0.81	0.84	0.81
ECOR38-b	0.75	0.62	0.85	0.96	0.74	0.82	0.76	0.75	0.74	0.75	0.78	0.81	0.64	0.72	0.64	0.82		0.79	0.78	0.73	0.74	0.74	0.70	0.80	0.75	0.77
ECOR57-b	0.81	0.68	0.80	0.79	0.83	0.84	0.74	0.79	0.71	0.72	0.75	0.82	0.70	0.77	0.65	0.76	0.79		0.86	0.71	0.77	0.71	0.68	0.82	0.75	0.76
ECOR67-b	0.83	0.76	0.83	0.78	0.85	0.97	0.79	0.78	0.74	0.75	0.80	0.84	0.74	0.80	0.71	0.80	0.78	0.86		0.75	0.79	0.73	0.67	0.86	0.81	0.78
ECOR72-b	0.76	0.66	0.84	0.73	0.73	0.81	0.95	0.83	0.59	0.70	0.78	0.84	0.73	0.72	0.67	0.79	0.73	0.71	0.75		0.82	0.58	0.67	0.78	0.81	0.76
GMB07-b	0.69	0.66	0.84	0.73	0.68	0.83	0.83	0.98	0.62	0.69	0.75	0.87	0.72	0.66	0.66	0.82	0.74	0.77	0.79	0.82		0.63	0.64	0.81	0.89	0.82
GMB23-b	0.72	0.73	0.65	0.72	0.73	0.76	0.61	0.63	0.99	0.69	0.67	0.69	0.72	0.70	0.77	0.69	0.74	0.71	0.73	0.58	0.63		0.60	0.69	0.68	0.65
GMB52-b	0.64	0.62	0.68	0.73	0.68	0.70	0.68	0.66	0.60	0.95	0.69	0.66	0.60	0.62	0.62	0.71	0.70	0.68	0.67	0.67	0.64	0.60		0.75	0.68	0.71
GMB76-b	0.79	0.72	0.85	0.81	0.81	0.90	0.81	0.81	0.69	0.82	0.93	0.83	0.70	0.77	0.67	0.81	0.80	0.82	0.86	0.78	0.81	0.69	0.75		0.80	0.78
GMB83-b	0.73	0.75	0.83	0.73	0.72	0.85	0.82	0.88	0.67	0.74	0.78	0.88	0.77	0.72	0.71	0.84	0.75	0.75	0.81	0.81	0.89	0.68	0.68	0.80		0.83
H1-1218-b	0.71	0.71	0.81	0.75	0.69	0.79	0.78	0.83	0.64	0.76	0.74	0.82	0.74	0.66	0.67	0.81	0.77	0.76	0.78	0.76	0.82	0.65	0.71	0.78	0.83	

Figure 4.2. Correlation matrix between Biolog GN2 profiles (OD_{600} values) of strain replicates after 24h at 37°C. Two replicates "-a" and "–b" were performed for 13 randomly chosen strains. Values of the table are Spearman correlation coefficients (r_S) between whole profiles. The heatmap indicates high (green) to low (white) values of r_S. The H1-1218 strain was analysed with the help of Stephanie Schüller (IFR).

Apart from 2 isolates (GMB83 and the O104:H4 outbreak strain H1-1218), the highest r_S coefficient for every strain was always with the corresponding experimental replicate (corresponding p-values were always inferior to 0.0001; data not shown). Correlation coefficients between experimental isolates ranged from 0.74 to 0.99, whereas correlation with other non-replicate strains ranged from 0.58 to 0.89. This indicates that reproducibility is overall very good. In the following, due to the large amount of strains tested (n=170), we performed one incubation per strain only, but

replicated once again if results indicative of a possible contamination (multiple positive reactions in uncommon *E. coli* substrates) were obtained.

4.2.1.3. Qualitative observation of respiration and necessity for an optical density threshold determination

The presence of purple colour is indicative of tetrazolium reduction and thus of the utilisation of the substrate as a sole C-source. We first visually assessed the plates after 24h incubation at 37°C to produce a binary output indicating the utilisation (purple colour is seen) or not (the well remains colourless) of the 95 C-sources for 170 tested strains (GMB and ECOR). On average, the 170 tested strains were able to grow on 44/95 C-sources, or 46.3% of all substrates available on a GN2 plate. The number of C-sources used by either ECOR or GMB was also around 43 to 44 C-sources (**Figure 4.3**), although the difference between the two collections was statistically significant (two-tailed t-test; t=2.296; p=0.0229).

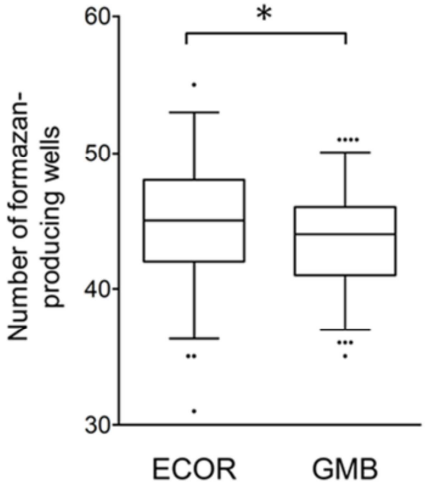


Figure 4.3. Number of C-sources from a Biolog GN2 plate utilised by ECOR (n=72) and GMB (n=98) strains after 24h incubation at 37°C (positive tests determined visually by assessing tetrazolium reduction). Boxplots represent the 25th to 75th percentile and outliers are represented by dots. The asterisk indicates statistical significance ($p<0.05$; see text).

Despite this coherent observation, the qualitative visual examination of test results is not very powerful and is very subjective for C-sources supporting slow growth, the purple colour being very faint. Not in possession of an Omnilog system, we chose to adopt the measurement of optical density using a classical 96-well plate spectrophotometer to get a quantitative dataset of C-source utilisation across our tested strains, as also performed in multiple studies (King et al., 2004; Ihssen et al., 2007; Maharjan et al., 2007). We found a very high correlation between tetrazolium reduction and OD_{600} general levels (data not shown), which led us to use OD_{600} measurements only and not rely on the visual inspection of the plates.

There is a similar problem of subjectivity when determining which C-source is used or not using OD_{600} data. Previous studies used arbitrarily decided thresholds for determining positive tests, such as $OD_{600}>0.2$ (Maharjan, Seeto et al. 2007). In order to reduce any subjective bias as much as possible, we developed a method to empirically determine a threshold for positive tests using OD_{600} and Biolog plates, with the help of Kate Kemsley (Bioinformatics & Statistics, IFR). First, a subset of 26 C-sources were identified as "non-utilised" from examination of the tetrazolium reduction and comparison between strains (**Table 4.2**)

Table 4.2. C-sources considered as "non-utilised" (utilised by less than 2% of all 170 strains) for empirical threshold determination.

α-Cyclodextrin	Succinamic Acid	Itaconic Acid	L-Pyroglutamic Acid
Glycogen	L-Alaninamide	α-Keto Butyric Acid	L-Threonine
Tween 40	L-Histidine	α-KetoValeric Acid	γ-Amino Butyric Acid
i-Erythritol	Hydroxy-L-Proline	Malonic Acid	Urocanic Acid
Xylitol	L-Leucine	Quinic Acid	Phenyethylamine
α-Hydroxybutyric Acid	L-Ornithine	Sebacic Acid	
γ-Hydroxybutyric Acid	L-Phenylalanine	2,3-Butanediol	

The distribution of the OD_{600} data from these sources is shown in **Figure 4.4A**. This figure is a good illustration of the need to setup an empirically defined threshold. All these C-sources are not utilised by any tested *E. coli* when the plates are visually assessed for tetrazolium reduction. Yet, the OD_{600} values range between 0 and 0.65 (**Figure 4.4A**), possibly because of inoculum carry-over or technical variability. If an arbitrary threshold was to be defined at OD_{600}>0.2 as suggested in previous studies (Maharjan, Seeto et al. 2007), C-sources such as the ones below that are not utilised by *E. coli* would appear positive.

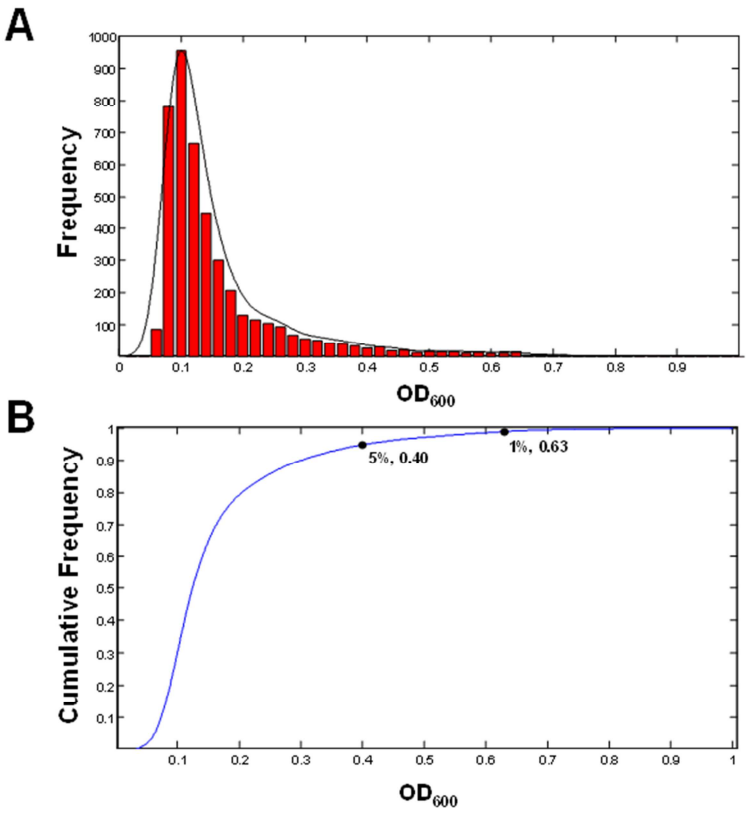


Figure 4.4. Empirical definition of a statistical threshold for positive carbon source utilisation; (A) Histogram (red columns) shows OD_{600} values across 26 non-utilised carbon sources; kernel density estimation of the probability density function is represented by the black line and its cumulative density function in panel (B), showing values for 5% and 1% tails. See text for details. This figure was produced with the help of E. Kate Kemsley (IFR).

In contrast to the approach reported by Sabarly et al 2011, we did not find it possible to adequately model the OD_{600} values using parsimonious Gaussian mixture models. Instead, kernel density estimation was used to obtain a non-parametric probability density function. The kernel bandwidth was optimised by cross-validation; a normal kernel function was employed. The thresholds demarcating the upper 5% and 1% tails of the distribution were found to be 0.40 and 0.63 respectively (**Figure 4.4B**). We have used this latter, more conservative threshold to pre-process the complete OD_{600} dataset, by setting all values less than 0.63 to zero; that is to say, only OD_{600} values higher than this threshold are taken as indicative of carbon source utilisation after 24 h incubation.

4.2.2. Variation in C-source utilisation by ECOR and GMB strains

4.2.2.1. General qualitative comparison of metabolic abilities

Using the threshold defined in section 4.2.1.3., we compared the number of positive tests between ECOR and GMB strains, as we did after visual assessment for tetrazolium reduction (**Figure 4.4**).

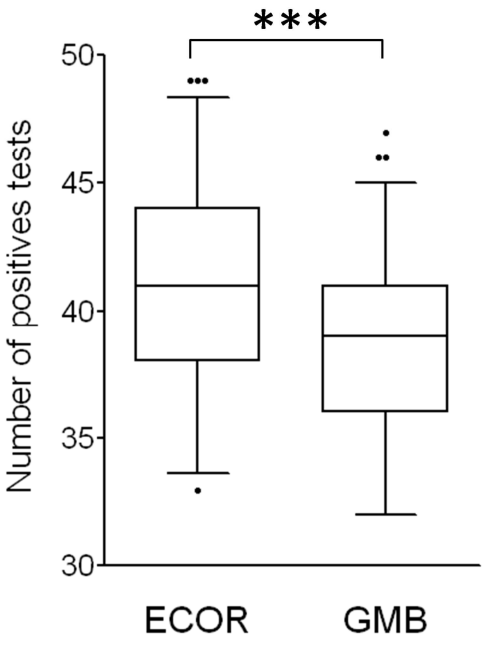


Figure 4.4. Number of C-sources from a Biolog GN2 plate utilised by ECOR (n=72) and GMB (n=98) strains after 24h incubation at 37°C (positive tests determined as tests with OD_{600}>0.63). Boxplots represent the 25th to 75th percentile and outliers are represented by dots. The asterisk indicates statistical significance ($p<0.0001$; see text).

The average numbers of C-sources used were 40.96 ± 0.5276 for ECOR strains and 38.54 ± 0.3656 for GMB strains, and their difference was very significant (two-tailed t-test; $t=3.890$; $p<0.0001$). The number of positive C-sources was slightly lower than what we had visually assessed earlier (**Figure 4.3**), which reflects the conservativeness of the approach we chose to adopt to define positive tests. We then compared how substrates from the different chemical guilds (**Table 3.1**) were utilised by ECOR and GMB strains (**Figure 4.5**).

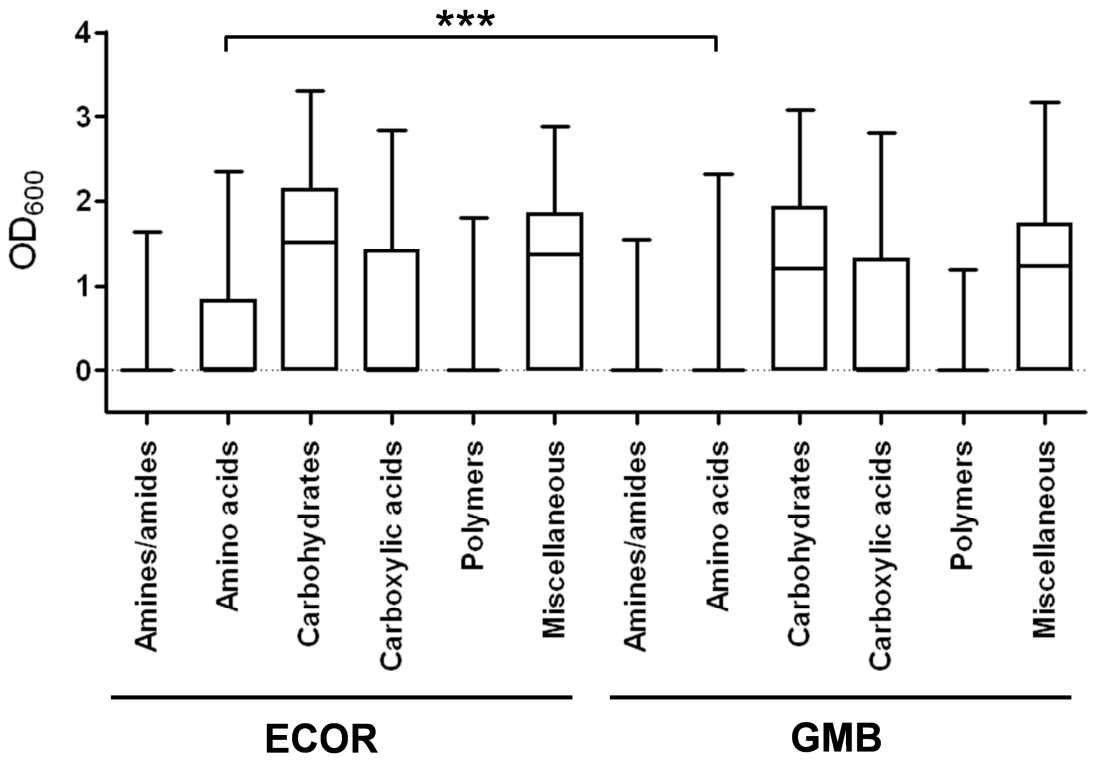


Figure 4.5. Utilisation of substrates from various chemical guilds by ECOR and GMB strains (see text and Table 3.1). Asterisks indicate statistical significance after a Dunn's post-hoc comparison test ($p<0.0001$).

A Kruskal-Wallis test expectedly confirmed the statistical differences between the utilisation of the different substrate guilds (statistic=2800; $p<0.0001$). More interestingly, a Dunn's post-hoc comparison test showed that significant differences between ECOR and GMB were restricted to the utilisation of amino acids ($p<0.0001$), whereas other guilds were not statistically different.

4.2.2.2. Identification of individual C-sources showing different utilisation patterns

To further the investigation on metabolic differences between ECOR and GMB strains, we compared utilisation at the individual C-source level (**Table 4.3**). We initially calculated the difference in percentage of strains from each collection able to use individual C-sources for growth at OD_{600}>0.63 after 24 h at 37°C (**Table 4.3**).

Table 4.3. C-sources showing more than 10% difference between the percentage of positive ECOR and GMB strains. Positive isolates grew at OD_{600}>0.63 after 24 h at 37°C (see text).

GN2 well	C-source	%ECOR	%GMB	Difference[a]
F09	L-Aspartic Acid	51.4%	7.1%	44.2%
G09	L-Serine	72.2%	39.8%	32.4%
G06	L-Proline	51.4%	20.4%	31.0%
G08	D-Serine	63.9%	36.7%	27.2%
E08	Propionic Acid	36.1%	9.2%	26.9%
H10	D,L-α-Glycerol Phosphate	93.1%	67.3%	25.7%
F05	D-Alanine	73.6%	52.0%	21.6%
F08	L-Asparagine	22.2%	3.1%	19.2%
B09	Lactulose	38.9%	20.4%	18.5%
E12	Succinic Acid	94.4%	76.5%	17.9%
F01	Bromosuccinic Acid	84.7%	67.3%	17.4%
H03	Uridine	93.1%	76.5%	16.5%
F10	L-Glutamic Acid	15.3%	0.0%	15.3%
A03	Dextrin	91.7%	77.6%	14.1%
F11	Glycyl-L-aspartic Acid	52.8%	38.8%	14.0%
C07	Sucrose	45.8%	68.4%	-22.5%
E01	*p*-Hydroxy-Phenylacetic Acid	43.1%	67.3%	-24.3%
C04	D-Raffinose	51.4%	76.5%	-25.1%

a. Positive values indicate C-sources preferentially used (by more than 10%) by ECOR strains; negative values indicate C-sources preferentially used (by more than 10%) by GMB strains.

There were more C-sources preferentially used by ECOR than by GMB strains. The overall biggest differences between the two collections were for amino acid utilisation

as C-source, as 8 amino acids out of 20 present on GN2 plates were preferentially used by ECOR strains (**Table 4.3**), possibly explaining the statistical difference between ECOR and GMB amino acid utilisation observed earlier (**Figure 4.5**). For example, more than half of all ECOR strains were able to use L-aspartic acid after 24 h at 37°C compared to slightly more than 7% of GMB only (44.2% difference). Similarly, generally around 30% more strains in ECOR were able to use L- and D-serine and L-proline (**Table 4.3**). Only 3 C-sources were preferentially used by GMB: sucrose, *p*-hydroxy-phenylacetic acid (*p*-HPA) and D-raffinose (**Table 4.3**).

To test if these differences were important in the overall metabolic profile variation between ECOR and GMB strains, we used multivariate analysis. The OD_{600} dataset was treated with partial least square discriminant analysis (PLS-DA) to look for evidence of grouping according to ECOR or GMB (**Figure 4.6**). PLS is a supervised modelling method that differs from the well-known PCA in the sense that the user *a priori* specifies the groupings to test (hence the supervision), in our case, ECOR and GMB. The consequence is that even a PLS plot resulting from random values will show groupings, which is why cross-validation is required to test the robustness of the specified groups.

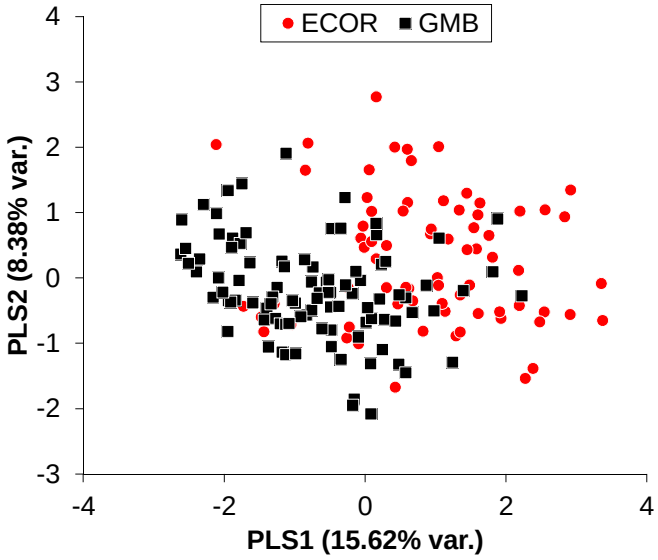


Figure 4.6. PLS-DA plot using ECOR and GMB as groups. The % in axes labels indicate how much of the total dataset variation is represented by the corresponding axes. This figure was created with the help of E. Kate Kemsley (IFR).

In our case, it was found that 2 PLS dimensions alone are sufficient to discriminate the collections with a cross-validation success rate of 81.8%. Adding a third PLS dimension increased the success rate to 83.5% which was the maximum discrimination ability that could be obtained (**Table 4.4**).

Table 4.4. Cross-validation success rates in PLS-DA using ECOR and GMB as groups for model dimensions up to 10. The maximum cross-validation success rate is highlighted and represents 3 PLS-DA dimensions. On Figure 4.6, we represented only 2 dimensions for clarity, which corresponded to a 81.8% success rate. This table was created with the help of E. Kate Kemsley (IFR).

PLS-DA model dimensions	By collection		Across all observations
	ECOR	GMB	% correct
	% correct		
1	75	78.6	77.1
2	80.6	82.7	81.8
3	**79.2**	**86.7**	**83.5**
4	77.8	82.7	80.6
5	76.4	83.7	80.6
6	76.4	82.7	80
7	76.4	83.7	80.6
8	72.2	82.7	78.2
9	70.8	83.7	78.2
10	68.1	81.6	75.9

This PLS-DA analysis confirms that ECOR and GMB strains form distinct groups and do not have the same Biolog GN2 profiles. In other words, isolates from secondary nonhost environments have different metabolic profiles than host isolates after 24 h incubation at 37°C. In order to get some ecological meaning from these trends, we performed nonparametric Mann-Whitney-Wilcoxon tests for each individual C-source to determine which were showing the most significant difference between ECOR and GMB (**Figure 4.7**, **Table 4.5**). This analysis varies from the result of **Table 4.3** in which we only compared proportions of strains from each collection. In **Figure 4.7** and **Table 4.5**, we show the output of a statistical comparison taking into account the levels of OD_{600} reached after 24 h incubation at 37°C. C-sources significantly associated with collection-associated variation are presented in **Figure 4.7**, along with the frequency distribution of OD_{600} of these C-sources.

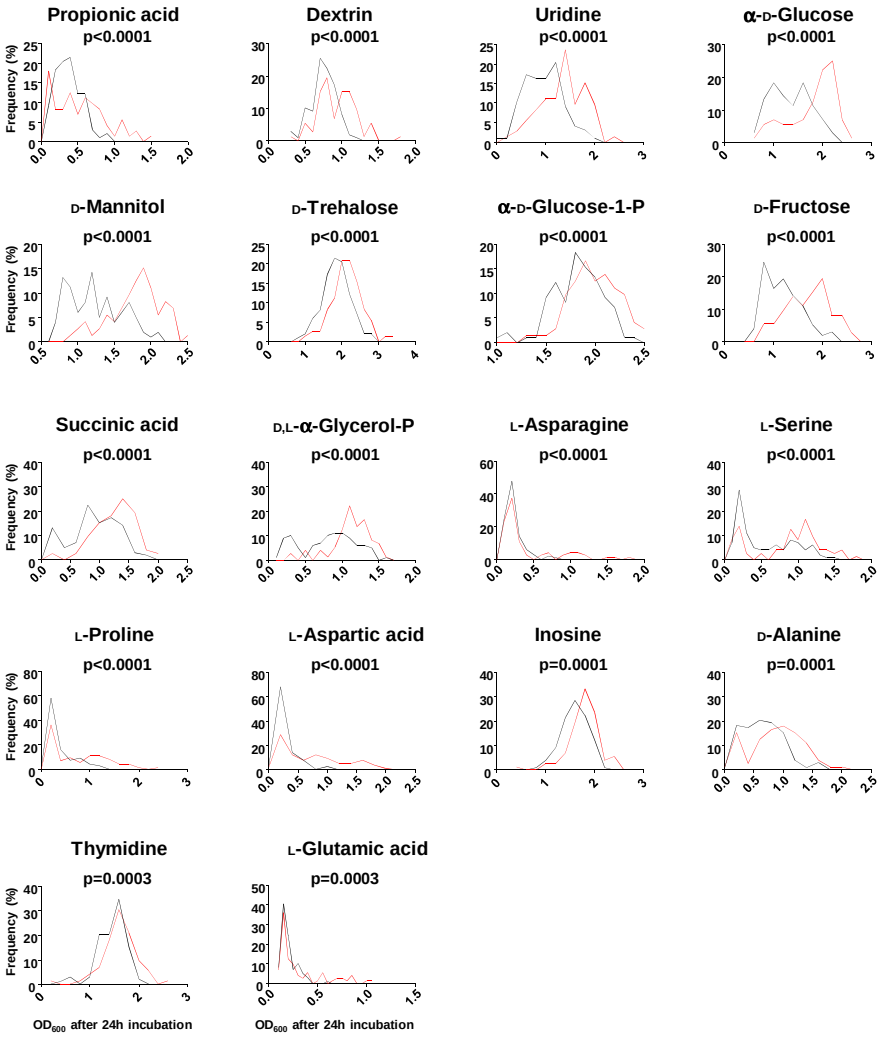


Figure 4.7. C-sources statistically associated with collection-associated variation in OD600 levels after 24 h of incubation at 37°C on Biolog GN2 plates. Red line is ECOR, black line is GMB. Indicated p-values are from individual Mann-Whitney-Wilcoxon tests for each C-source with a Bonferroni correction.

Strikingly, the C-sources found to cause the most variation between ECOR and GMB were not completely overlapping the results of **Table 4.3**, indicating that the levels of OD_{600} reached after 24 h at 37°C are important factors in defining the difference between the two collections. These C-sources found to differ in levels of utilisation were always generally used at higher levels by ECOR strains than GMB strains after 24 h, as suggested by the frequency plots on **Figure 4.7**. This observation was

particularly true for α-D-glucose, D-mannitol, D-trehalose, D-fructose, and succinic acid.

Table 4.5. C-sources found to show most variation between ECOR and GMB and their utilisation abilities. This table was created in collaboration with E. Kate Kemsley (IFR).

GN2 well	C-source	Utilisation (OD_{600}>0.63)			MWW p-value	Median OD_{600}	
		%ECOR	%GMB	Diff.		ECOR	GMB
A03	Dextrin	91.7%	77.6%	14.1%	<0.0001	0.98	0.76
B02	D-Fructose	100.0%	98.0%	2.0%	<0.0001	1.79	1.13
B06	α-D-Glucose	100.0%	99.0%	1.0%	<0.0001	1.95	1.37
B11	D-Mannitol	100.0%	100.0%	0.0%	<0.0001	1.86	1.18
C08	D-Trehalose	100.0%	100.0%	0.0%	<0.0001	2.14	1.86
E08	Propionic Acid	36.1%	9.2%	26.9%	<0.0001	0.54	0.38
E12	Succinic Acid	94.4%	76.5%	17.9%	<0.0001	1.32	0.90
F09	L-Aspartic Acid	51.4%	7.1%	44.2%	<0.0001	0.62	0.22
G06	L-Proline	51.4%	20.4%	31.0%	<0.0001	0.70	0.23
G09	L-Serine	72.2%	39.8%	32.4%	<0.0001	0.96	0.49
H03	Uridine	93.1%	76.5%	16.5%	<0.0001	1.39	0.98
H10	D,L-α-Glycerol Phosphate	93.1%	67.3%	25.7%	<0.0001	1.13	0.88
H11	α-D-Glucose-1-Phosphate	100.0%	100.0%	0.0%	<0.0001	2.03	1.85
F08	L-Asparagine	22.2%	3.1%	19.2%	<0.0001	0.22	0.20
F05	D-Alanine	73.6%	52.0%	21.6%	0.0001	0.96	0.58
H02	Inosine	98.6%	100.0%	-1.4%	0.0001	1.80	1.60
F10	L-Glutamic Acid	15.3%	0.0%	15.3%	0.0003	0.21	0.18
H04	Thymidine	98.6%	98.0%	0.7%	0.0003	1.61	1.51

Interestingly, many C-sources showing the most variation between ECOR and GMB often used by most, if not all the strains (**Table 4.5**) suggest that the uptake of C-sources is differentially affected in the two collections rather than the presence or absence of metabolic genes in one and not the other. This suggestion is consistent with the observation that most of the C-sources in **Figure 4.7** and **Table 4.5** are common substrates for growth (α-D-glucose, D-mannitol, D-trehalose, D-fructose).

The lower levels of growth on common C-sources exhibited by GMB strains can be caused by a slower metabolism. To confirm this observation, we selected a subset of 5 ECOR and 10 GMB strains that we grew using Biolog GN2 plates at 24 h, 48 h and 72 h (**Figure 4.8**).

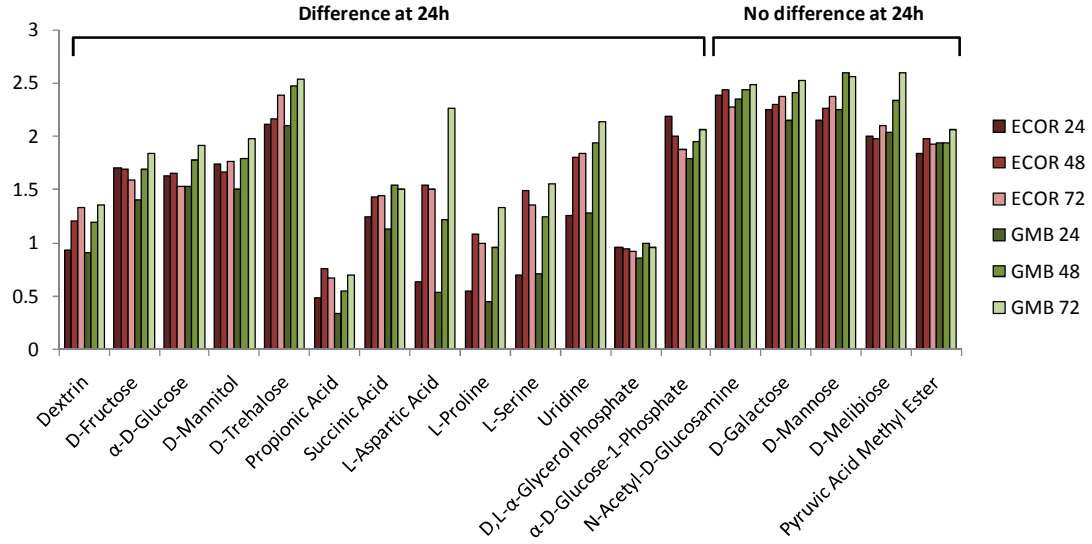


Figure 4.8. Comparison of average growth levels reached after 24 h, 48 h and 72 h by 5 ECOR strains and 10 GMB strains on Biolog GN2 plates on selected C-sources. Red bars represent averages of 5 ECOR results after 3 different times of incubation (24, 48 and 72 h). Green bars represent averages of 15 GMB results after 3 different times of incubation (24, 48 and 72 h). Thirteen C-sources showing the most difference between ECOR and GMB at 24 h were included, as well as 5 C-sources showing no difference between ECOR and GMB at 24 h.

Although the 13 C-sources we examined in this experiment all had lower levels at 24 h in the GMB strains, they showed the same levels of utilisation as ECOR strains after 72 h. The only statistically significant difference we found was for the utilisation of L-aspartic acid by GMB and ECOR strains at 72 h (two-sided t-test; $t=3.277$; $p<0.05$), which reached very high levels in GMB comparatively to ECOR. However, we did not find any statistical significance in the rest of the data, even for the 24 h time point,

most probably because of the low number of strains used. In any case, the general trend was that although growth levels at 24 h were lower for GMB strains, in almost all the cases, they were slightly higher than ECOR at 72 h, indicating that the low metabolic abilities observed for GMB strains at 24 h are possibly caused by a slower assimilation and utilisation, and not by lower efficiency of metabolic pathways. To be certain of this, a reproduction of this experiment with a higher number of strains grown in Biolog GN2 for longer than 24 h would be necessary.

In this section, we introduced a novel way to use OD_{600} data from Biolog experiments to compare metabolic profiles of large numbers of isolates and showed phenotypical differences between plant and faecal isolates of *E. coli*. We observed after 24 h of incubation at 37°C that there were differences in the preferential utilisation of amino acids by ECOR and of sucrose, D-raffinose and the aromatic compound *p*-hydroxyphenylacetic acid by GMB. We also showed that most of the C-sources explaining the biggest difference between the two collections of isolates were common C-sources used by all strains, possibly because of the different speed of uptake or metabolism. To investigate this last point further, we checked if there were any links between genetic content and the Biolog metabolic profiles using comparative genomic hybridisation (CGH).

4.2.2.3. Genetic association with metabolic data using CGH

In order to associate genetic information with observed metabolic phenotypes, we conducted association tests using Genespring built-in MWW tests with Bonferroni corrections. The Bonferroni correction is the most stringent multiple testing correction

method, which is suitable for microarray studies as given the experimental variation intrinsic to DNA hybridisation, weakly significant false positives can be common. Using these settings on the data produced by the genomes of 21 GMB and 20 ECOR strains hybridised on ShEcoliO157 microarrays (see Methods section 2.5), we sought to associate Biolog data to genomic information (**Figure 4.9**).

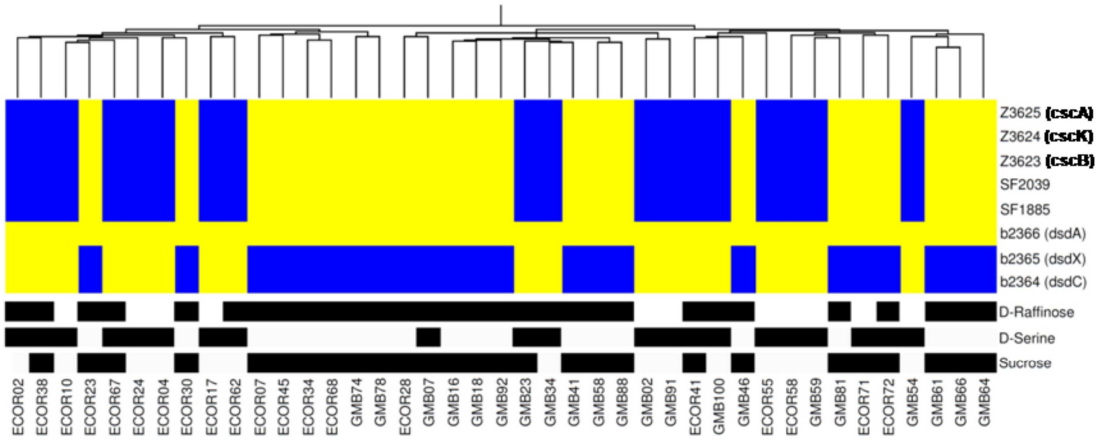


Figure 4.9. CGH data for genes statistically associated with sucrose, raffinose and D-serine Biolog utilisation patterns. Blue denotes absent genes, yellow denotes present gene. The dendrogram above the gene information panel corresponds to clustering according to the Pearson correlation on the whole hybridised genomes. Below genetic information, black squares denote positive Biolog test for the corresponding C-source and white denotes a negative test. The *dsdA* gene was not statistically associated but was added to the figure to show the whole *dsd* operon (see text).

We could associate the D-serine utilisation patterns with the well-described *csc-dsd* locus (**Figure 4.9**). Although *dsdA* was not found by association, it was added to confirm that the observed *csc* gene insertions were in accordance with the previously known mechanism of insertion into the *dsd* locus (Jahreis et al., 2002), which results in the deletion of *dsdX* and *dsdC* leaving *dsdA* intact. Surprisingly, not a single gene was statistically associated with sucrose utilisation after Bonferroni correction, although the inverted relationship between these two metabolisms is clear (**Figure**

4.9) and has been extensively documented (Alaeddinoglu and Charles, 1979; Jahreis et al., 2002). When a different false discovery rate correction method was used (Benjamini and Hochberg FDR correction), which in our experience is less stringent than the Bonferroni method, the same genes as those associated with the D-serine utilisation patterns were identified. This difference in association stringency could be explained by the fact that *E. coli* has 2 different pathways for sucrose metabolism (Reid and Abratt, 2005), but only one to metabolise D-serine. This explanation is supported by the observation that the inverted sucrose/D-serine phenotype was not systematic as some isolates could metabolise both sucrose and D-serine (**Figure 4.9**), indicating that some of our dsd^+ strains can still use sucrose via another pathway, including plasmid-borne systems. Cross-utilisation between sucrose and D-raffinose has also been previously documented (Arr et al., 1970). However, D-raffinose utilisation was only weakly and not statistically significantly associated with the *csc-dsd* locus. Also associated with sucrose and D-serine were 2 prophage genes (SF2039, a putative Q antiterminator of prophage and SF1885, a putative tail component. It has been documented that *E. coli* O157:H7 possessed phage-like elements in the proximity of the *csc-dsd* locus, which seemed to be conserved in other strains due to the proximity with a prophage insertion hotspot (Moritz and Welch, 2006).

p-HPA is one of the compounds preferentially used as a sole carbon source by GMB over ECOR strains (**Table 4.3**). Only when no multiple testing corrections were applied, the genes most associated with *p*-HPA metabolism were expectedly found to be among the *hpa* catabolic locus (data not shown).This lack of strong association suggests that there may be additional regulatory factors to consider for explaining aromatic compound degradation pathways, such as the possible degradation by other

proteins encoded by genes from other aromatic compounds clusters (Diaz et al. 2001) that are not present on the microarrays. Another explanation on why genetic associations with metabolic phenotypes are weak may be that the chosen growth conditions in Biolog plates are not optimal to reflect the corresponding metabolic genetic content of our strains. Indeed in 5 strains, there was a discrepancy between the metabolic observation and the presence of *hpa* genes (GMB02, 18 and ECOR17 showed no *p-HPA* degradation *in vitro* despite harbouring *hpa* genes; and GMB23 and ECOR71 showed the phenotype but did not harbour *hpa* genes; data not shown). Accepting that we use a conservative approach in our Biolog experiment (some phenotypes may not be detected after 24 h even if the strain is actually able to use the corresponding C-sources) it is understandable that genetic and phenotypic information are not perfectly matched. It may also just be that we did not include enough samples in the CGH study (41 strains out of 178 strains, or only about 23% of all strains in our collection are represented). Lastly, the *E. coli* pangenome has been estimated to include an estimation of 18,000 gene families (Lukjancenko et al. 2010), which obviously are not represented on a ShEcoliO157 array.

We tested all other C-sources present on a Biolog plate but not a single other one showed any association with genetic information. The biggest limitation of the approach detailed in this section is that association can only be made when we observed phenotypical variation. This suggests that within the boundaries of our study, the *csc-dsd* locus was the most highly variable and possibly correlated genetic cluster with Biolog GN2 carbon source utilisation patterns.

4.2.3. Phylogenetic distribution of metabolic abilities

A key consideration in any microbial population study is the link between phylogeny and ecology. Being informed about the phylogenetic relationships between strains (Chapter 3 of this thesis) and some of their phenotypes, we were interested in connecting both by examining whether strains from different phylogenetic backgrounds have different ecological strategies, as monitored by their metabolic profiles. This type of association has rarely been shown in *E. coli* (Sankar et al. 2009; Sabarly et al. 2011). We used the same PLS-DA statistical pipeline as the one we used to examine variability across collections (**Figure 4.6**, **Table 4.4**), this time using phylogroups as a potential discriminating factor (**Figure 4.10**, **Table 4.6**).

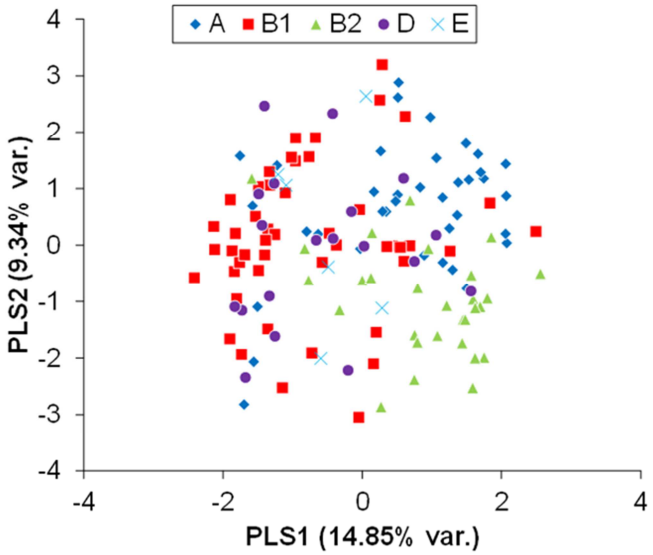


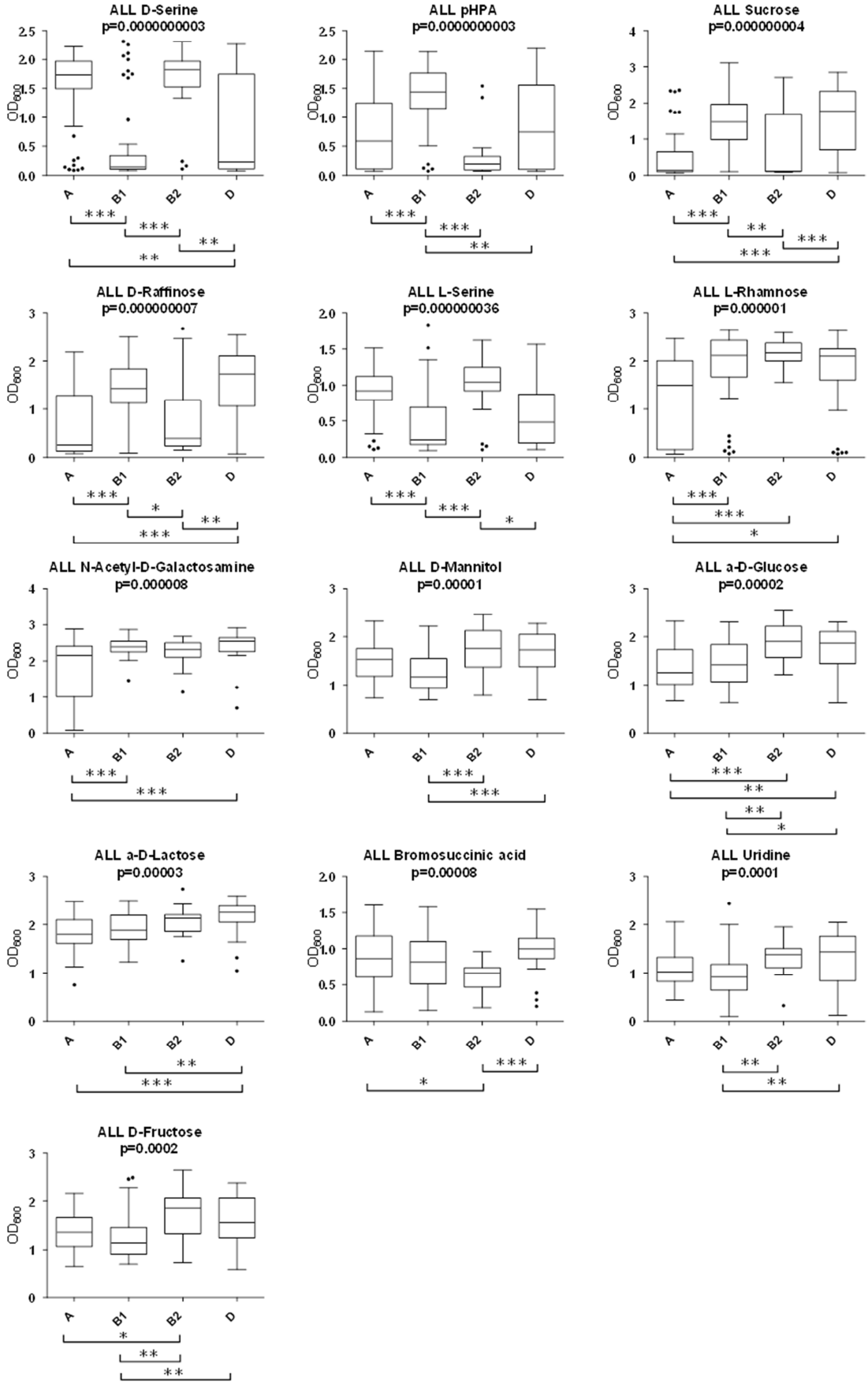
Figure 4.10. PLS-DA plot using phylogroups A, B1, B2, D and E as groupings to test. Blue diamonds are strains from phylogroup A, red squares from B1, green triangles from B2, purple circles from D and cyan crosses from E. The % in axes labels indicate how much of the total dataset variation is represented by the corresponding axes. The corresponding cross-validation calculations are shown in Table 4.6. This figure was created with the help of E. Kate Kemsley (IFR).

Visually, from the PLS-DA plot (**Figure 4.10**), it was strikingly obvious that profiles from phylogroup A and B1 were very different. Phylogroup B2 strains also seemed to discriminate, and it was less obvious for phylogroup D strains. As expected, 2 PLS dimensions alone were sufficient to discriminate the collections with a cross-validation success rate of 62.5%. The maximum successful discrimination was obtained using 4 PLS dimensions, for a cross-validation rate of about 75% (**Table 4.4**).

Table 4.6. Cross-validation success rates in PLS-DA using phylogroups as groupings for model dimensions up to 10. The maximum cross-validation success rate (74.4%) is highlighted and represents 7 PLS-DA dimensions. On Figure 4.6, we represented only 2 dimensions for clarity, which corresponded to a 62.5% success rate. Done with the help of E. Kate Kemsley (IFR)

PLS-DA model dimensions	A	B1	B2	D	All
1	63.3	65.5	31.8	35.5	55.0
2	77.6	69.0	63.6	45.2	62.5
3	75.5	67.2	77.3	58.1	63.1
4	71.4	77.6	81.8	61.3	67.5
5	75.5	77.6	77.3	58.1	70.6
6	75.5	77.6	72.7	58.1	72.5
7	73.5	75.9	72.7	51.6	74.4
8	71.4	79.3	72.7	51.6	73.8
9	71.4	81.0	77.3	51.6	73.1
10	69.4	81.0	72.7	54.8	70.6

These calculations indicate that whole Biolog profiles can be discriminated according to the phylogroups of the corresponding isolates. In other words, strains from different phylogenetic backgrounds have different metabolic abilities. In order to identify these differences, we performed individual Kruskal-Wallis tests for each C-source using thresholded OD_{600} data. We found 13 C-sources whose utilisation distribution was not random across phylogroups (**Figure 4.11**).


ALL D-Serine
p=0.0000000003
ALL pHPA
p=0.0000000003
ALL Sucrose
p=0.000000004
ALL D-Raffinose
p=0.000000007
ALL L-Serine
p=0.000000036
ALL L-Rhamnose
p=0.000001
ALL N-Acetyl-D-Galactosamine
p=0.000008
ALL D-Mannitol
p=0.00001
ALL a-D-Glucose
p=0.00002
ALL a-D-Lactose
p=0.00003
ALL Bromosuccinic acid
p=0.00008
ALL Uridine
p=0.0001
ALL D-Fructose
p=0.0002
OD600
A
B1
B2
D

← Figure 4.11. Phylogenetically distributed C-source utilisation patterns in our tested E. coli collection (n=170). ECOR and GMB strains are mixed. C-sources are ordered according to the p-values after individual Kruskal-Wallis tests. The p-values have been adjusted using the Bonferroni correction. Boxplots represent the distribution of OD_{600} values for each phylogroup. Phylogroups E and F were excluded from the analysis because of their low sample size. The asterisks represent statistically significant differences between groups after a Dunn's post-hoc comparison test and the number of asterisks represent different significance thresholds (* for $p<0.05$; ** for $p<0.001$; *** for $p<0.0001$).

Visually, the differences in C-source utilisation distribution after 24 h of incubation at 37°C for the most significantly associated C-sources (D-serine, *p*-HPA, sucrose, D-raffinose and L-serine) is strikingly clear, and is reflected by a very low Kruskal-Wallis p-values. D-serine utilisation, the most significantly associated with a phylogroup-dependent variation was generally used by phylogroup A and B2 strains and not by phylogroup B1 and D strains (**Figure 4.11**). As expected, the D-serine and sucrose/raffinose distributions are inverted, with strains from phylogroups B1 and D using sucrose and raffinose in clear majority. Surprisingly, *p*-HPA is also very strongly associated with a non-random phylogenetic distribution, and follows the same pattern as for sucrose utilisation. This observation interestingly suggests that the acquisition and maintenance of accessory metabolic genes (such as the sucrose and aromatic compounds metabolic clusters) is not uniform within bacterial species but is linked to specific phylogenetic histories.

It was shown previously that *p*-HPA is generally not used by phylogroup B2 strains (Sabarly et al. 2011), as we also observed. However, for sample homogeneity reasons, Sabarly et al. (2011) grouped strains A and B1 together, missing the differences

between these phylogroups. When groups were compared between them, it is actually B1 that was statistically different ($p<0.001$) from all others (**Figure 4.11**).

Differences in the utilisation patterns of other associated C-sources did not follow the same trends. L-rhamnose and N-acetyl-D-galactosamine showed a very strong difference in utilisation by strains from phylogroup A. Common C-sources, such as glucose, fructose and lactose also showed phylogenetically-distributed utilisation patterns (**Figure 4.11**). Interestingly, we observed a generally great difference in utilisation patterns between phylogroup A and B1 strains, which is unexpected as these phylogroups are believed to have recently diverged, and thus share higher degrees of functional similarity, which we show here is not the case.

4.2.4. Conclusive remarks on host vs. nonhost metabolic variability

Based on the observations of the metabolic study presented in section 4.2., we can conclude that there are two different levels of metabolic differences between nonhost and host isolates of *E. coli*. First, the metabolism of common "core" C-sources (used by all strains of *E. coli*) mostly occurs at lower speeds in nonhost strains compared to host strains. Secondly, the possession of specific genetic information (the laterally acquired sucrose and aromatic compound metabolic clusters) is also significantly playing a role in differentiating nonhost from host strains.

Intra-species differences in levels of nutrient acquisition have previously been linked to variations in the self-preservation and nutritional competence (SPANC) balance (Ferenci, 2005). It is believed that within bacterial populations such as *E. coli*, there is

a trade-off between the acquisition of nutrients and resistance to stressful conditions, as a correlation between those two traits has been observed (De Paepe et al., 2011) and even linked to intracellular RpoS concentration (Ferenci, 2005; De Paepe et al., 2011; Ferenci et al., 2011). It sounds indeed ecologically plausible that GMB strains have adapted to better resist stressful conditions such as those found in nonhost environments, especially if they transferred from multiple environments such as water and soil before being isolated from plants. However, although we have a possible indirect indication of it with the lower nutrient acquisition abilities we observed, we did not test for stress resistance in this work. This information could confirm or not the hypothesis that the SPANC balance has an important ecological role in nonhost adaptation by environmental *E. coli* strains isolated from plants.

As a major difference from faecal isolates of the ECOR collection, GMB plant isolates were observed to be better metabolisers of sucrose, a carbohydrates used as an energy transport molecule found in great abundance in plant vascular tissues and raffinose, a plant-associated trisaccharide sharing metabolic requirements with sucrose. It is worth noting that in *E. coli*, sucrose utilisation has been described to be inversely linked to D-serine utilisation, because of a genetic insertion of the *csc* genes of sucrose metabolism in the *dsd* (D-serine utilisation) metabolic cluster (Jahreis et al., 2002). In our Biolog dataset, the sucrose profiles were indeed most significantly correlated with the D-raffinose profiles (Spearman correlation; $r_S=0.6068$, $p\approx10^{-18}$) and the D-serine profiles ($r_S=–0.4652$, $p\approx10^{-10}$). This high correlation observed between raffinose and sucrose profiles is consistent with a shared metabolic pathway, as it indicates that many sucrose-positive or -negative strains are also correspondingly raffinose-positive or -negative. Plant isolates were also observed to be better

metabolisers of *p*-HPA, an aromatic compound generally found in abundance in soils (Diaz et al., 2001). These differences have obvious ecological meaning for these agricultural isolates and allow us to get a clearer picture on the signatures of nonhost association in plant-associated *E. coli*.

4.3. Colonisation-associated phenotypes

To further our phenotypic characterisation and comparative analysis of plant and faecal isolates of *E. coli*, we performed assays to test for phenotypes that have been reported to generally enhance the colonisation of new environments: biofilm formation, flagellar motility and siderophore production.

4.3.1. Comparison of biofilm formation by plant and host isolates and their motility

Indeed, components of the extracellular matrix, the most important factor in biofilm formation by *Enterobacteriaceae*, have been linked to an increased persistence and attachment on plants (see Introduction section 1.3.2.1) and an increased survival to dessication, an important type of stress in the phyllosphere. We used the crystal violet assay (see Methods section 2.7.1) to quantify biofilm formation among our isolates. Briefly, this *in vitro* assay relies on the irreversible staining of bacteria with crystal violet. When forming a biofilm on polystyrene surfaces (i.e., at the surface of microtitre plates wells, bacteria become firmly attached and cannot be washed away, as opposed to planktonic bacteria, and can thus be quantified). We used conditions previously known (Landini et al. 2006) to induce higher biofilm production in *E. coli*:

lower temperature (28°C-30°C) and growth in CFA medium (see Methods section 2.7.1 and Introduction section 1.3.2.1).

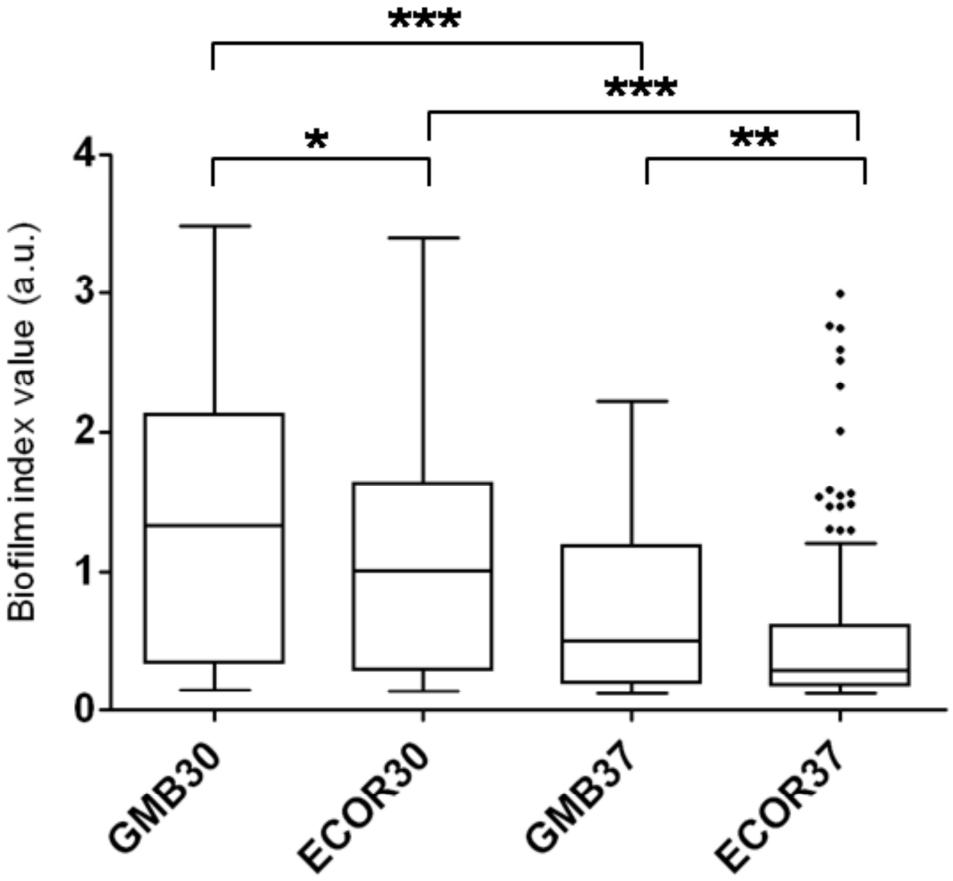


Figure 4.12. Biofilm formation by ECOR (n=72) and GMB (n=106) strains after 72 h of incubation at 30°C and 37°C. Asterisks indicate statistical significance after a Dunn's post-hoc comparison after a Kruskal-Wallis test. "30" or "37" after GMB or ECOR label denotes the temperature at which biofilm formation was examined. The number of asterisks reflects the significance threshold, with *: $p<0.05$; **: $p<0.001$ and ***: $p<0.0001$.

Expectedly, all strains formed very significantly less biofilm at 37°C than at 30°C (**Figure 4.12**). Also, GMB strains formed statistically more biofilm than ECOR strains, and this difference was more significant at 37°C than at 30°C. This result indicates that GMB strains are able to form more biofilms *in vitro* than ECOR strains.

As we mentioned above, the advantage of biofilm formation has been suggested for the establishment of *E. coli* and *Salmonella* in the phyllosphere. First described as an advantage in aquatic ecosystems, it is also believed that biofilm formation could help bacteria resist desiccation in stressful living conditions such as those that can arise in agricultural conditions on plants, but also in soils. In that sense, it is not surprising to see that *E. coli* strains isolated from nonhost environments tend to be more efficient at producing an extracellular matrix than faecal isolates. The possibility that biofilm formation by *E. coli* provides a selective advantage during nonhost life has already been hypothesised based on the dissection of the genetic regulation affecting biosynthetic genes related to the extracellular matrix (Landini et al. 2006). Indeed, the capsule synthesis master regulator CsgD and the biosynthetic *csg* operons encoding curli, the major proteinaceous fimbriae of the enterobacterial extracellular matrix are upregulated at low temperature, low osmolarity and during stationary phase of growth, which would tend to correspond more to extra- than intra-host life conditions (Landini et al. 2006). Our observation that nonhost-associated strains produce more biofilm than host-associated strains also contributes to the hypothesis that biofilms are ecologically important in *E. coli* for nonhost lifestyles and persistence (see Introduction section 1.3.2.1).

Flagellar motility has been linked to an increased colonisation of rocket salad and spinach leaves by pathogenic *E. coli* and of basil and lettuce leaves by *Salmonella enterica* (see Introduction section 1.3.2.1). We then examined the frequency of motility among ECOR and GMB strains. We qualitatively assessed the ability of strains to swim in commercial motility agar after stabbing of an overnight culture (**Figure 4.13**).

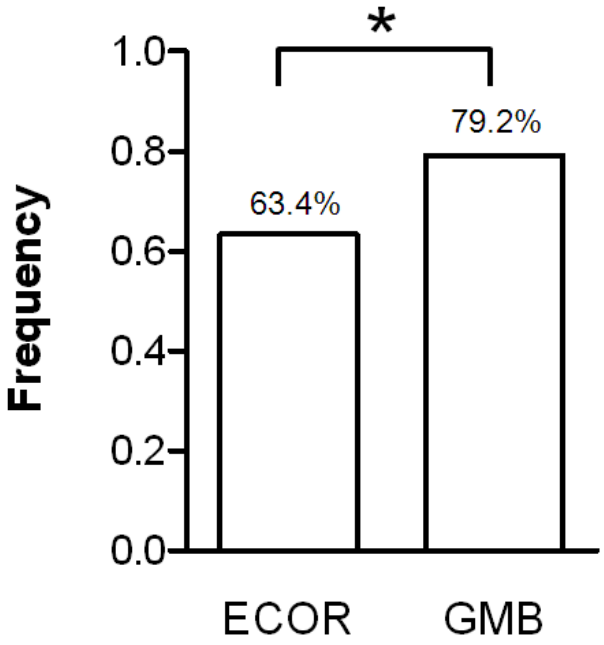


Figure 4.13. Flagellar motility frequencies at 37°C for ECOR (n=71) and GMB (n=101) strains. The asterisk indicates statistically significant frequencies ($p<0.05$).

GMB strains were found to be more motile at 37°C than ECOR strains (GMB=80/101, 79.2%; ECOR=45/71, 63.4%; Two-sided χ^2 test; $\chi^2=5.259$; $p=0.0218$). In the light of the previously published literature on the importance of flagella for plant colonisation, this observation is ecologically significant and as for the increased biofilm formation, may reflect enrichment for ecologically relevant plant colonisation mechanisms in plant isolates of *E. coli*.

4.3.2. Effect of plant auxin-derivatives on biofilm formation by plant isolates of *E. coli*

If and how plants interact with non-pathogenic immigrating bacteria or even with seemingly neutral (i.e. non-interacting) microbial communities living on or next to them remains poorly understood. It is believed that the synthesis of plant hormones, some of which are volatile, can affect phytosphere microbial communities and

immigrating bacteria. Two major hormones of plants, jasmonic acid (JA) and salicylic acid (SA) have been shown to effect on the structure and diversity of epi- and endophytic microbial communities of *Arabidopsis thaliana* (Kniskern et al., 2007). It was shown that impairing the *Arabidopsis* SA-mediated pathways reduced endophytic diversity, whereas the lack of the JA-mediated defenses pathway increased the diversity of epiphytic bacteria (Kniskern, Traw et al. 2007). Regarding immigrating bacteria, it has been shown *in vitro* that the plant auxin derivatives had an impact on *E. coli in vitro* cultures. The addition of the plant auxin indole-3-acetic acid (IAA) to growing cultures of *E. coli* K-12 increased their resistance *in vitro* to various stresses and antibiotics, and slightly but significantly increased their biofilm formation (Bianco et al., 2006). More recently, contradicting results have been produced, as IAA and more importantly 3-indoleacetonitrile (IAN) could very significantly decrease (and not increase) *in vitro* biofilm formation of *E. coli* O157:H7 (Lee et al., 2011). From a comparison of the protocols used in these two studies, it would seem that the difference is more attributable to strain-specific effects rather than a concentration or a temperature effect.

It is then very interesting to examine the effect of plant auxin derivatives on biofilm formation in large number of strains, including the ecologically relevant GMB collection. In this section, we first examined the toxicity of auxin IAA on our *E. coli* strains (**Figure 4.14**). We then added non-toxic concentrations of plant auxin IAA and derivative IAN to bacterial cultures during Kolter assays to monitor their effect on biofilm formation (**Figure 4.15**).

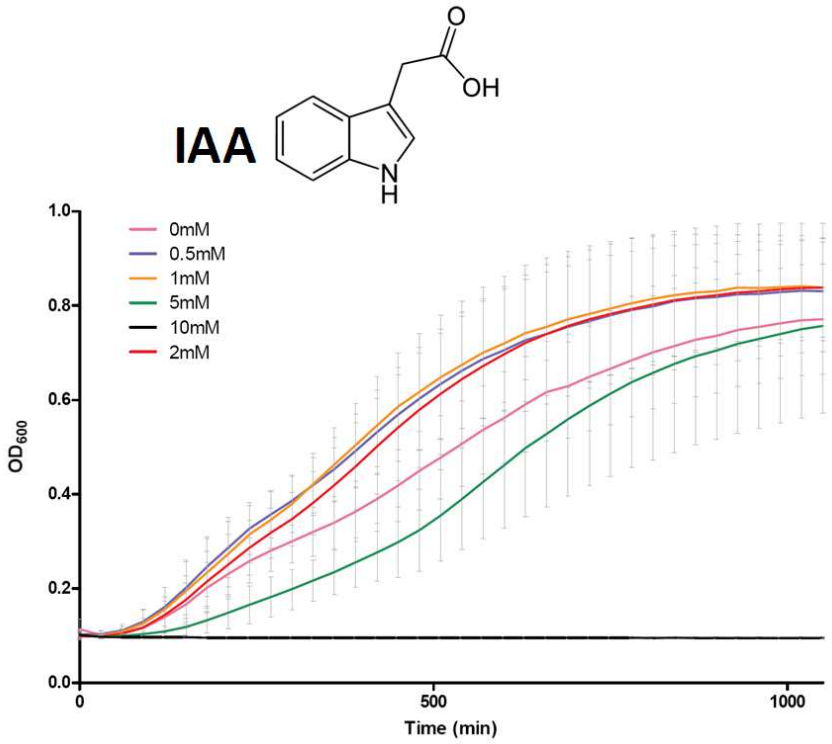


Figure 4.14. *E. coli* growth in LB in the presence of various concentrations of plant auxin indole-3-acetic acid (IAA). The growth curves show the averages of 16 *E. coli* strains (ECOR and GMB). Error bars denotes standard deviation from the average.

Unfortunately, we did not include IAN in our toxicity tests, but the response to various concentrations in two other auxins, IAA and a closely related molecule, phenylacetic acid (PAA), was striking between them (data not shown). For both IAA and PAA, the maximum concentration with no negative effect on growth was 2mM (data not shown). Concentrations below 2 mM seemed to improve growth, compared to the control (medium alone). This is maybe caused by the IAA (or PAA) starting to be metabolised by the strains at these concentrations (*E. coli* can harbour *paa* genes, involved in PAA, and possibly IAA metabolism). Bianco et al. (2006) cautiously used 0.5 mM in their assays whereas Lee et al. (2011) showed that IAN was not toxic to *E. coli*, even at concentrations as high as 150 µg/mL. Based on this information and our toxicity experiment, we decided to use 2 mM IAA and IAN in the following biofilm

experiments. We added 2 mM IAA or IAN to bacterial cultures and let them grow statically for 72 h at 30°C or 37°C before performing a Kolter assay on the plates.

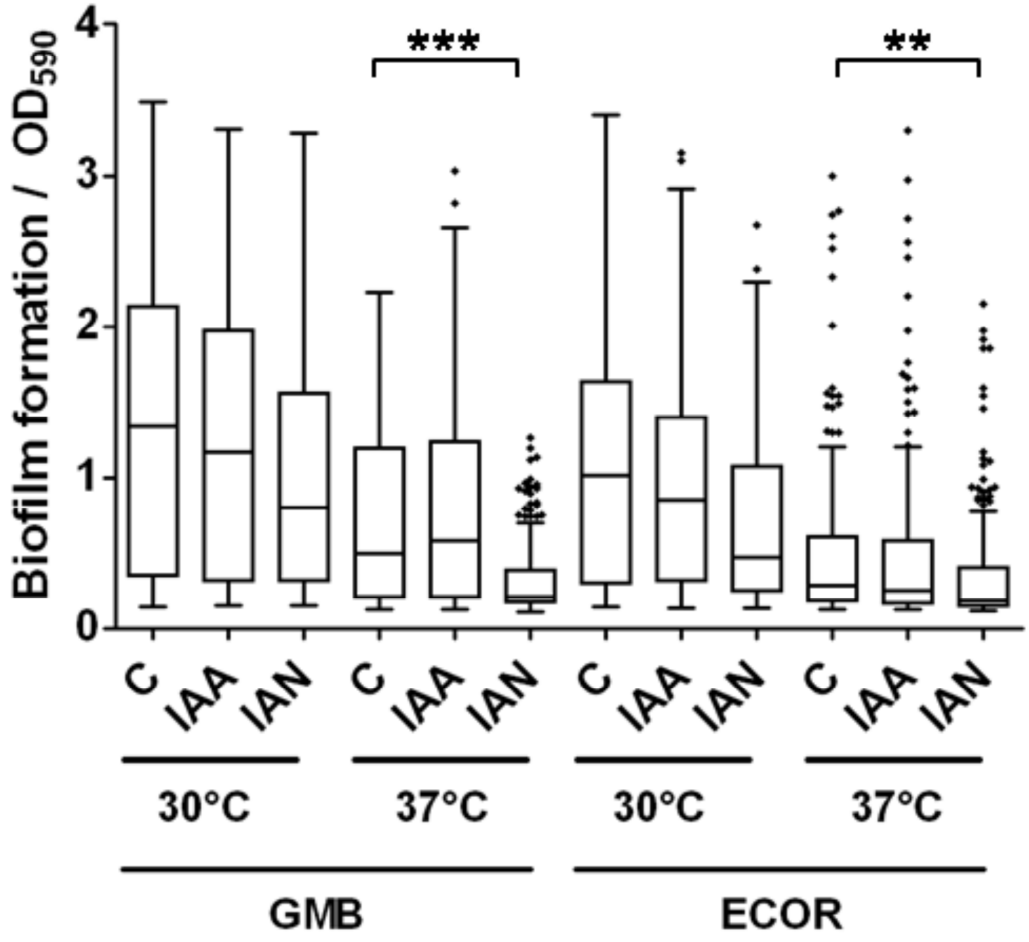


Figure 4.15. Effect of 2 mM IAA and 2mM IAN on biofilm formation by ECOR (n=72) and GMB (n=106) strains after 72h of incubation at 30°C and 37°C. Asterisks indicate statistical significance compared to control values based on a Dunn's post-hoc comparison after a Kruskal-Wallis test. The number of asterisks reflects the significance threshold, with 2 asterisks corresponding to $p<0.001$ and 3 asterisks, $p<0.0001$.

The effect of 2 mM IAA and 2mM IAN on biofilm formation was not statistically different from the control for either GMB or ECOR strains at 30°C. However, IAN (but not IAA) had a very significant effect on GMB ($p<0.0001$) and ECOR ($p<0.001$)

biofilm formation at 37°C. The disparity of IAN effect at 37°C compared to 30°C can be compared to the previously published effect of IAN on *E. coli* O157:H7 at 37°C, which was very clear and much stronger than with IAA (Lee et al. 2011). As mentioned above, results from previously published literature are different, but may be attributed to the examination of single strains, which is always open to variable results, and to differences in protocol. Indeed, Bianco et al. (2006) compared adhesion of cells to polysterene after 20 h at room temperature, whereas Lee et al. (2011) compared adhesion after 24 h at 37°C. As we showed using our large collection of strains, the response at these two temperatures is very different and likely to be strain-specific.

4.3.3. Siderophore production

Iron in its ferric form (Fe^{3+}) is essential for many biochemical processes but is very insoluble. Its bioavailability is a major limiting factor for bacterial growth in natural environments and bacteria such as *E. coli* have evolved mechanisms to scavenge iron molecules in order to maintain their vital iron intracellular concentration between 10^{-7} to 10^{-5} M (Garenaux et al., 2011). It has been estimated that in order to survive and multiply within hosts, pathogenic strains of *E. coli* require 10^5 to 10^6 Fe^{3+} ions each generation. *E. coli* strains are able to synthesise up to 4 distinct siderophore molecules, named after genus names of *Enterobacteriaceae*: enterobactin, salmochelin, yersiniabactin and aerobactin (Garenaux, Caza et al. 2011). For *E. coli*, siderophores have been mainly described in pathogens as virulence factors and are suggested as target molecules for antibacterial compounds to limit pathogenic growth. Similar approaches have been undertaken in soil for rhizospheric bacteria and plant

pathogens (Jurkevitch et al., 1992; Cornelis, 2010; Diallo et al., 2011). Indeed, it is believed that iron is also very limiting for growth in nonhost environments such as soil and is therefore the object of intense competition by resident communities. Nevertheless, it was shown using biosensors that the availability of ferric iron was spatially heterogeneous on plant leaves, and that iron starvation at the bacterial scale on the phyllosphere was probably not uniform, with subsequently varying levels of competition (Joyner and Lindow 2000). Siderophore production by bacteria seem to be important for both host and nonhost interactions. To our knowledge, no published study has addressed the role of siderophore production during *E. coli* or *S. enterica* interactions with nonhost environments or even plants, so we included it in our comparative phenotypical analyses.

4.3.3.1. Siderophore production differences in plant and faecal isolates

We used chrome azurol S (CAS)-based solid medium to assess siderophore production *in vitro* (see Methods section 2.7.3) and compared the distribution of siderophore production between ECOR and GMB strains (**Figure 4.16**).

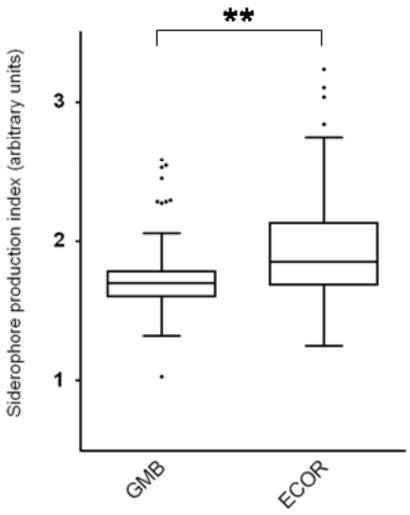


Figure 4.16. Siderophore production by ECOR (n=72) and GMB (n=106) strains. The asterisk denotes statistical significance based on a Bonferroni post-hoc comparison after a Welch's t-test (see text).

GMB strains produced significantly less siderophores (unpaired t-test with Welch's correction; t=3.09, p=0.0025) than ECOR strains *in vitro* on CAS indicator agar medium (**Figure 4.16**), with the top 11 (out of 178) siderophore producers being ECOR strains (data not shown).

4.3.3.2. Siderophore production by pathogenic strains of E. coli

Siderophore production in *E. coli* has been described in the literature as a virulence factor, notably for ExPEC (including UPEC and related APEC) (Caza et al., 2008; Wiles et al., 2008; Garenaux et al., 2011). It has been shown that siderophore production is important for host colonisation during the pathogenic process, and interestingly, we can observe that host-associated ECOR strains produce more siderophores than nonhost GMB strains (**Figure 4.16**). If siderophore production is a trait strongly associated with host-associated lifestyles, there is a possibility that siderophore production by pathogenic isolates is higher than GMB. It then becomes interesting to compare siderophore production between plant isolates and pathogenic isolates to see if the decreased utilisation by GMB fits this ecological hypothesis. We

performed a CAS indicator assay using APEC and Shiga-toxin mutants of *E. coli* O157:H7 (**Figure 4.17**).

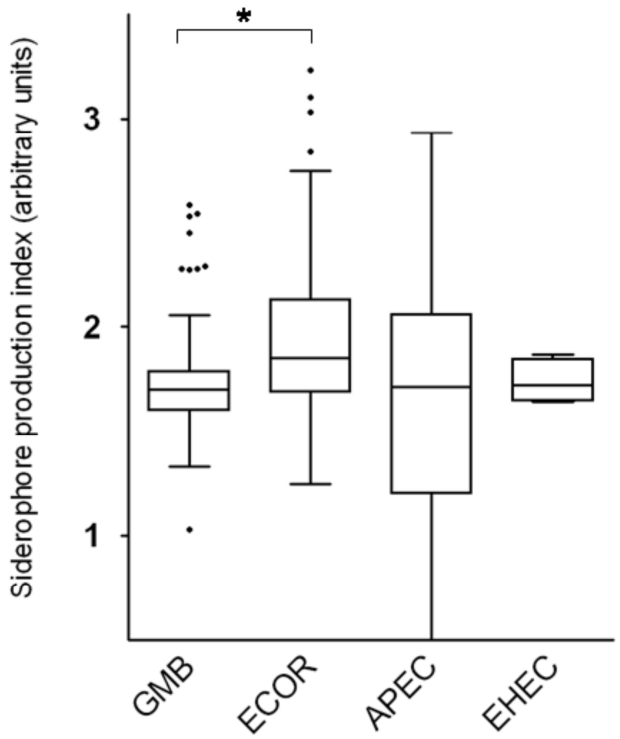


Figure 4.17. Siderophore production by ECOR (n=72), GMB (n=106), APEC (n=10) and EHEC (n=4) strains. The asterisk denotes statistical significance based on a Bonferroni post-hoc comparison after a one-way ANOVA (see text).

We did not detect any statistical difference between siderophore production of GMB or ECOR and pathogenic isolates (**Figure 4.17**). The only statistically significant difference was between GMB and ECOR as shown before (**Figure 4.16**). This observation suggests that ecological differences in siderophore production within the *E. coli* species are not great, possibly confirming their vital role in bacterial persistence and growth.

It is possible that the observed difference between host and nonhost strains is caused by our way of assaying siderophore production. We grew strains at 37°C, a temperature which is maybe not very conducive of siderophore production. *E. coli* has been documented to produce up to 4 different siderophore molecules (Garenaux, Caza

et al. 2011). It is therefore possible that only some of these siderophores are expressed and exported at 37°C on CAS indicator plates, therefore making plausible ecological explanations difficult to formulate, although it could indicate that GMB isolates are less able to produce siderophores at host temperature. Unfortunately, we did not perform siderophore production tests at lower temperature, but we have developed a method to detect the presence of the 4 different siderophore biosynthesis operons in a subset of *E. coli* strains. Using this method, we can try to correlate the siderophore production phenotype with the presence or absence of specific siderophores biosynthesis operons.

4.3.3.3. Development of a multiplex PCR tool for the detection of siderophore biosynthesis genes in E. coli

It is interesting to know if the differences in siderophore production indices observed in the previous sections are correlated with the ability to synthesise more types of siderophores, or just higher amounts of individual siderophores. To address this, we developed a multiplex PCR method to detect in 4 electrophoresis gel lanes 15 siderophore biosynthetic genes encoding for 4 different siderophore molecules (see Methods section 2.7.3). We used this method to test for the presence of siderophore genes in 28 ECOR and GMB strains classified as low (average siderophore index=0.686), medium (average siderophore index=1.783) and high (average siderophore index=2.825) siderophore producers (**Figure 4.18**).

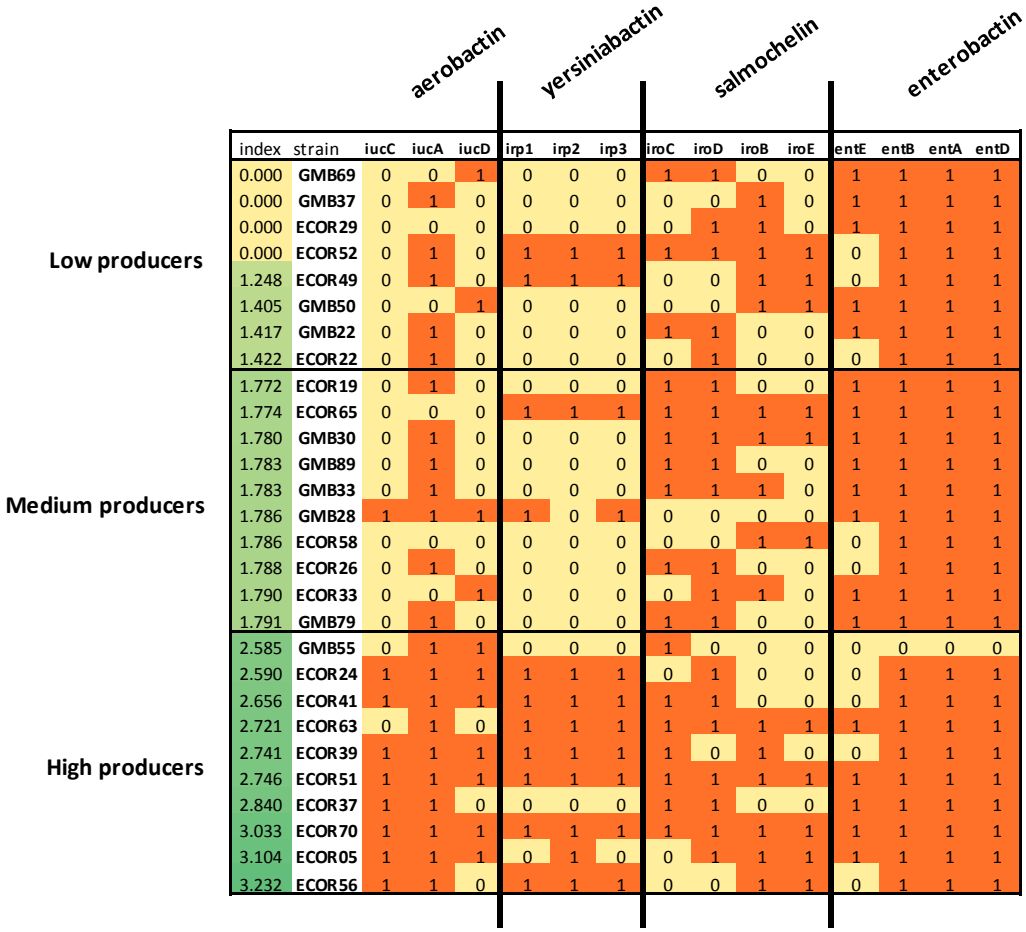


Figure 4.18. Graphical representation of band detection after a multiplex PCR detection test for siderophore biosynthetic genes. Orange colour indicates presence of a band, yellow colour indicates absence of band. The "index" column indicates values of the siderophore production index as measured on CAS indicator plates.

Visually, it seemed that high producers yielded more bands (i.e. harboured more siderophore biosynthetic genes). Using this multiplex PCR method, there were strains with only single genes detected and not the full set of genes of the operons. We observed that multiplexing sometimes prevents the amplification of target sequences, especially if these targets are closely located on the chromosome (data not shown). We also observed that enterobactin genes are almost always present in all tested *E. coli* strains, even those that did not grow or produce a halo on CAS agar (**Figure**

4.18), suggesting that the presence of genes alone is only one factor for siderophore production and that gene regulation is important. To further this observation, we calculated the average number of bands detected for low, medium and high siderophore producers (**Figure 4.19**).

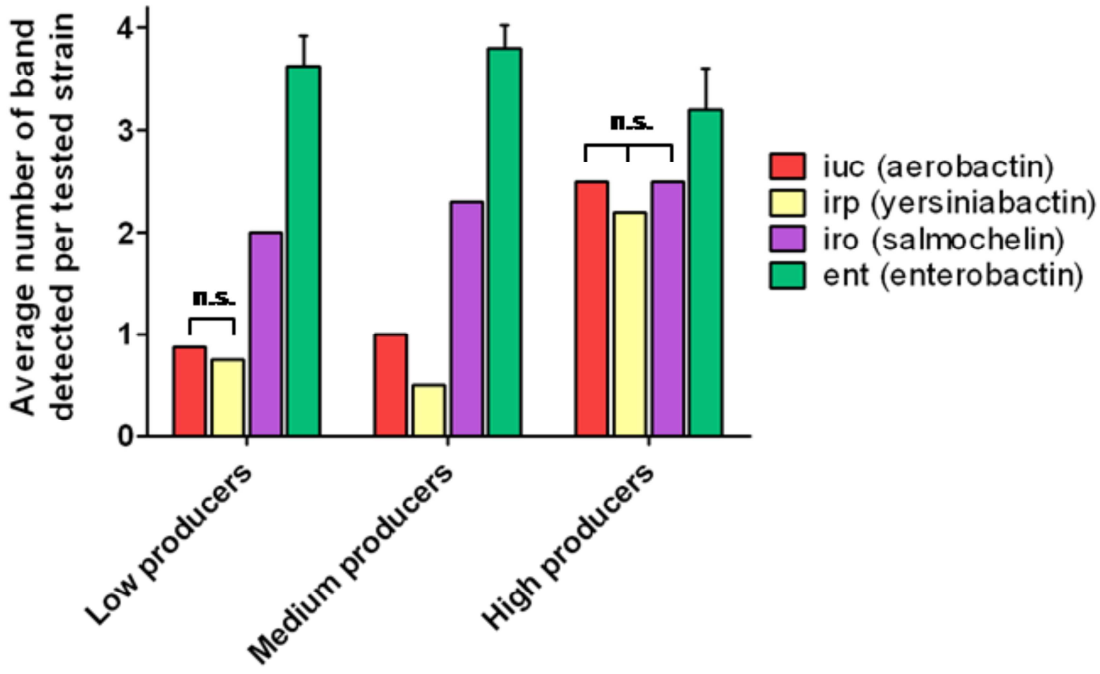


Figure 4.19. Average number of bands for low (n=8), medium (n=10) and high (n=10) siderophore-producing strains after multiplex PCR detection tests. As most of the values were significantly different after a two-way ANOVA, we only indicated the non-significant differences (n.s.) on the figure for clarity purposes.

Enterobactin and salmochelin genes were present in all types of siderophore producers. There was a notable increase in detection of genes for aerobactin and yersiniabactin in high producers compared to low and medium producers, suggesting that the ability to produce a diversity of siderophore and not necessarily a high amount of each is perhaps also causing the large halo phenotype on CAS indicator agar. In any case, there seems to be a link between phenotype on CAS agar and the

number of siderophore biosynthetic genes detected by PCR in the corresponding strains. However, for some siderophores like enterobactin, gene regulation seems to be involved, as strains harbouring a full *ent* operon can still not produce siderophore on CAS agar under our tested conditions.

4.3.3.1. Possible ecological explanations for the observed difference in siderophore production between host and nonhost isolates

The competition for iron seems to be ubiquitous, occurring in various environments. It is therefore surprising to see significant differences in siderophore production between *E. coli* strains of ecologically different origins of isolation. It seems very clear that ECOR produces more siderophores under our experimental conditions, but why? Is the competition for iron more important for *E. coli* in the gastrointestinal tract than in nonhost environments? We propose in this section two different but not exclusive explanations.

The first explanation is that the difference we observe could be caused by an experimental bias. We performed our CAS assay at 37°C on agar plates and obtained a large variation in siderophore production between our strains. However siderophore production, costly for a cell to maintain, may be under complex gene regulation processes and therefore our assay perhaps did not encompass all optimal conditions for siderophore production in *E. coli*. It seems to be the case, as we found full enterobactin synthesis operons in strains that were not even growing on CAS (meaning that they were not able to scavenge iron from the CAS-iron complexes, likely because they did not synthesise any siderophore molecule). If those

enterobactin genes are functional and not cryptic, they were therefore not expressed on CAS agar at 37°C, but it then does not mean that the strain is unable to produce siderophores.

Despite gene regulation considerations, it is also possible that the difference in siderophore production by GMB and ECOR strains reflects a biological reality. We show in this section that GMB isolates are producing less siderophores than ECOR strains at a temperature relevant for host colonisation. Additionally, siderophore production could be maintained at a high level when competition for iron is high. Because the possibility of "cheating" may be higher among phylogenetically related species than for distant bacteria, cross-feeding on excreted siderophores (as most of them are) is theoretically higher when the concentration of members of the same species is also high. *E. coli* most probably do not live in close contact with other *E. coli* cells in nonhost environments as it does in host gastrointestinal tracts. Alternatively, this mechanism could be species-independent, with physical proximity and population density being a key factor. Therefore, the need to synthetise high levels of siderophores in nonhost environments is not as crucial as in the gut, where most secreted siderophores can be more easily taken up by neighbouring "cheater" cells. One can then imagine that when competition for ferric iron is not strong, the optimal number of siderophore molecules synthesised for an *E. coli* cell is the minimum number that procures the minimum concentration of iron required for growth. However, in the gut, in order to grow, a cell must probably synthesise more siderophores for the same concentration of scavenged iron. In other words, the selective pressure to maintain high siderophore production could be caused by life in

environments where iron is limited, and where intra-species competition is very strong.

As an argument supporting this hypothesis, it has been shown that rhizospheric bacteria able to synthesise low-affinity siderophores were no longer able to grow if purified high-affinity siderophores were added to the medium (Joshi et al., 2006; Joshi et al., 2008). This growth inhibition was only relieved by the addition of iron to the medium (Joshi, Archana et al. 2006). This study suggests that in competitive environments, species producing low-affinity siderophores can be outcompeted by species producing higher-affinity siderophores. Interestingly, in stable communities and based on this study, one cannot help to wonder whether siderophore producers evolved to an optimal affinity level corresponding to their living environment. In that case, it would not be surprising to observe strains from the same species but from different environments having different capacities to scavenge iron as a result of competitive interactions with their ecological communities. These hypotheses are of course subject to experimental verification, but if proven true, they could provide an ecological explanation of the difference in siderophore production by GMB and ECOR strains.

4.4. Ecological ranking and the "plant association index"

4.4.1. The purpose of ecological ranking of isolates

Based on our observations from a collection of *E. coli* isolates from plants, we were able to dissect specificities and infer possible signatures or traits linked to the nonhost

environment where those strains were isolated. The “typical” plant-associated strain would be from phylogroup B1, motile, forming biofilm, having relatively low nutritional abilities after 24 h of growth on various C-sources including common ones, being a low siderophore-producer and a metaboliser of sucrose and the aromatic compound *p*-HPA.

4.4.1. Selection of phenotypes to rank and their phylogenetic distribution

In this section, we selected 5 phenotypic observations that we believe are ecologically relevant for *E. coli* plant persistence, and that we found to be associated with plant isolates of *E. coli* in our comparative analyses. We ranked individual strains based on their phenotypic values of biofilm formation, siderophore production, growth on sucrose and *p*-HPA and finally on their average OD_{600} on the 13 C-sources found to be variable between ECOR and GMB and shown in **Figure 4.7**. We showed in this section that some metabolic traits could be very significantly phylogenetically distributed, which prompted us to decompose the distributions of the phenotypes we chose to define a so called “plant association index”, or “PAi” according to phylogroups (**Figure 4.20**).

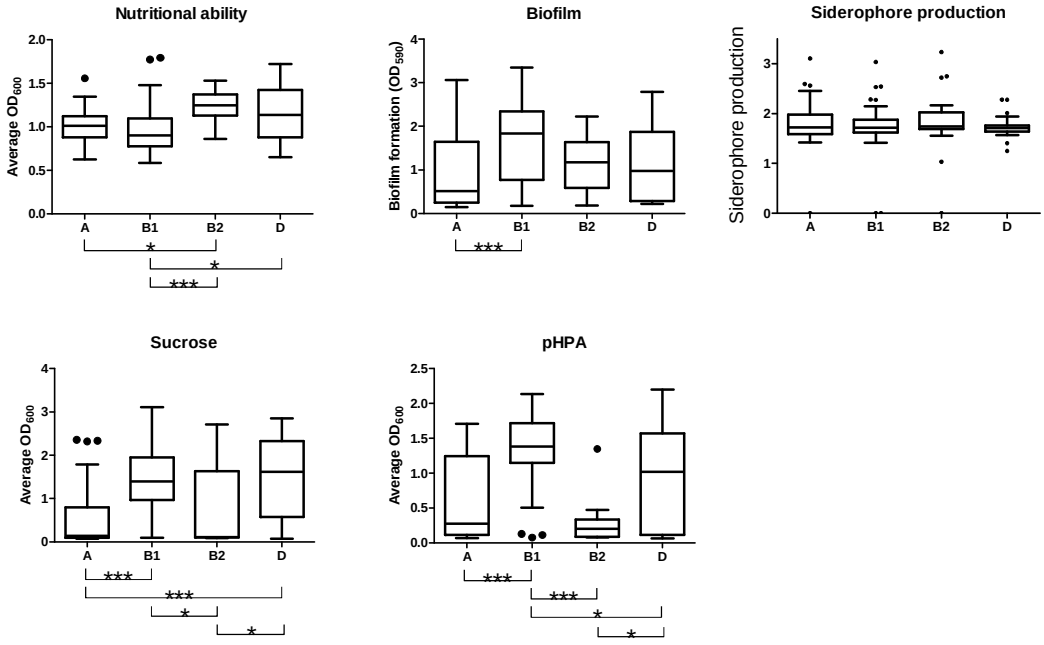


Figure 4.20. Phylogenetic distribution of 5 phenotypic factors used in the calculation of the "plant association" index (PAi). Asterisks indicate statistical significance based on a Dunn's post-hoc comparison after a Kruskal-Wallis test. The number of asterisks indicates the significance threshold with *, $p<0.05$; **, $p<0.001$ and ***, $p<0.0001$. ECOR and GMB strains are combined.

As mentioned before, sucrose and *p*-HPA metabolism were very significantly distributed according to phylogroups, and along the same trends: positive utilisation by phylogroup B1 and D strains, and negative utilisation by phylogroup A and B2 strains (**Figure 4.20, Figure 4.21**). Interestingly, nutritional abilities (the OD_{600} average on C-sources significantly discriminating between ECOR and GMB) were also distributed non-randomly across phylogroups, with major differences between phylogroups A/B1 and B2/D (**Figure 4.20, Figure 4.21**). A Kruskal-Wallis test found a non-random association with phylogroups for biofilm formation ($p=0.0005$), caused by a very significant difference between phylogroup A and B1 distribution of biofilm formation. The distribution of siderophore production however, was not phylogenetically distributed ($p=0.439$).

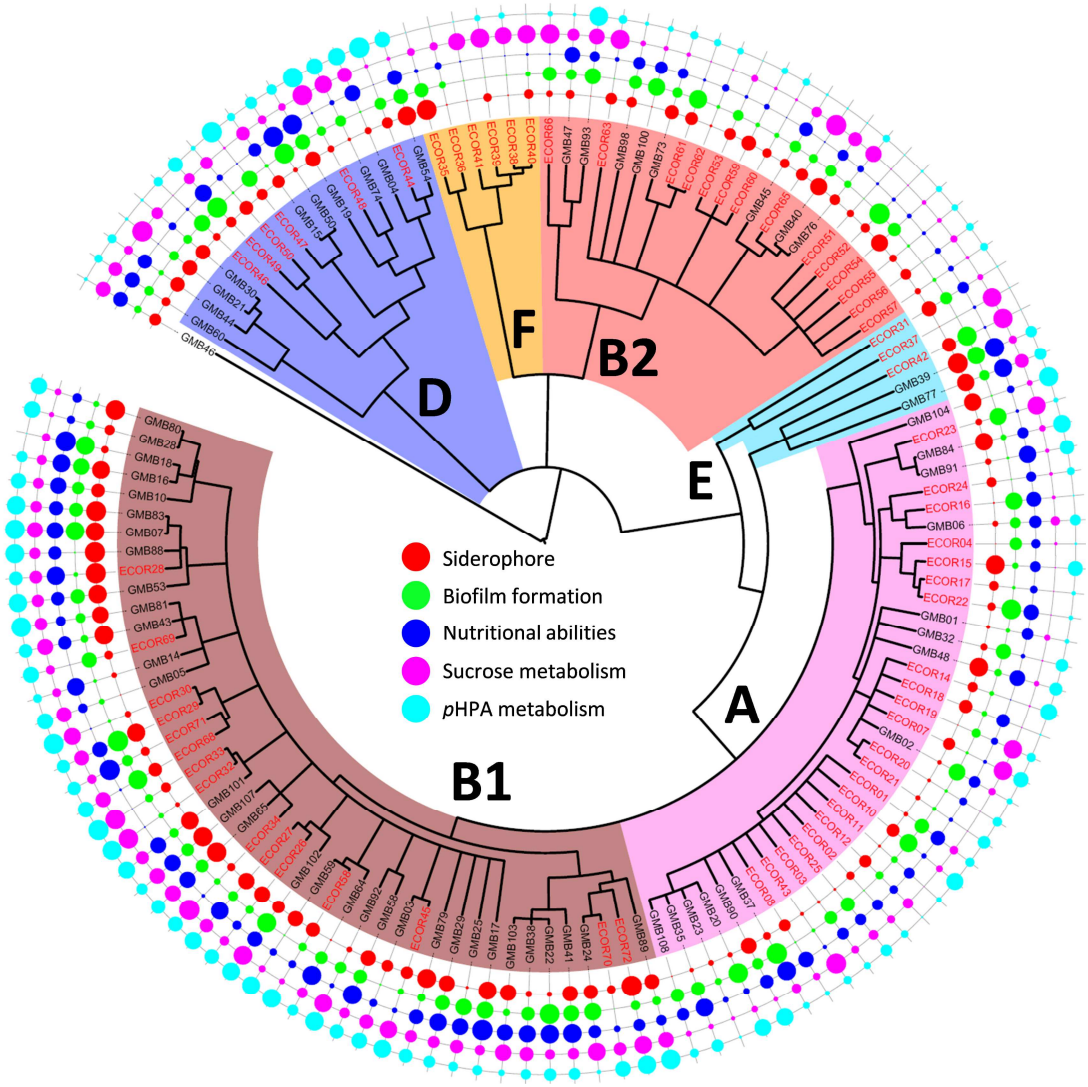


Figure 4.21. Distribution of selected phenotypic ranks on the ClonalFrame phylogenetic tree of ECOR and GMB isolates. The diameter of circles is proportional to the ranking, with a large diameter associated with a high ranking and a low diameter to a low ranking. Colour of the circles reflects the selected phenotype as indicated in the legend.

It was visually clear in **Figure 4.21** that these phenotypes have a tendency to distribute differentially according to phylogroups (except siderophore production). Very small circles, representing a low ranking for the chosen phenotypes, tend to cluster in phylogroup B2, F and to some extent in A, while large circles were

observed in phylogroup B1 and D. Siderophore production (red circles) was however more evenly distributed, consistent with the observation in **Figure 4.20**.

4.4.2. Combination of phenotype ranks and calculation of the PAi

In an effort to quantify and compare the potential for plant association among *E. coli* strains, we generated a plant association index (PAi) for each strain based on the traits discriminating GMB from ECOR as shown before: low siderophore production, high biofilm formation, low nutritional abilities, high metabolism on sucrose and *p*-HPA. The PAi was calculated as the sum of the phenotypical rank of each trait (**Figure 4.22**) and decomposed according to the strain phylogeny (**Figures 4.23 Figure 4.24**).

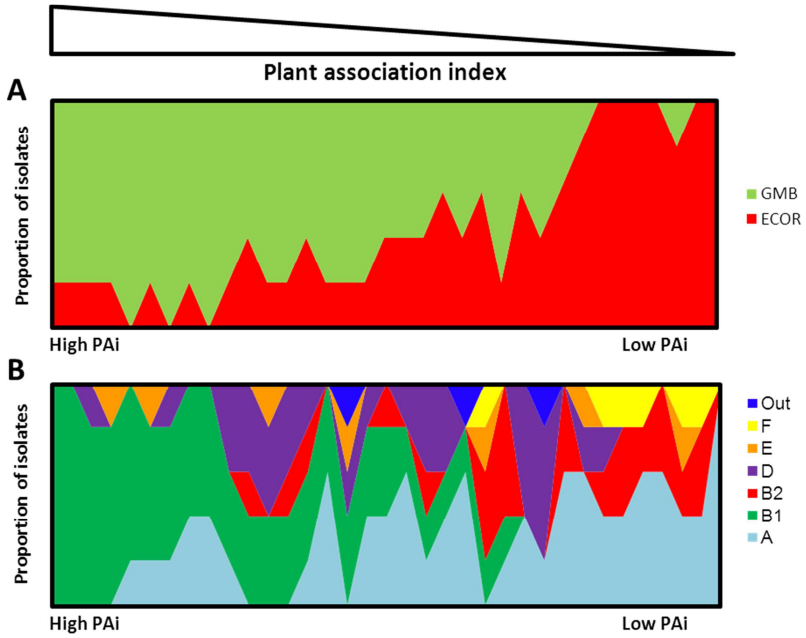


Figure 4.22. Frequency plot showing plant association index (PAi) order according to collection (A) or phylogroup (B). ECOR and GMB strains are combined in panel B. "Out" represent strains from *Escherichia sp.* Clade-I

As expected, the PAi was higher for GMB than for ECOR strains (**Figure 4.22A**). We also observed that the PAi was generally high for phylogroup B1 isolates and low for phylogroups A and B2 isolates (**Figure 4.22B**). This is also expected, as we showed in the previous sections that metabolic data such as sucrose and *p*-HPA, which accounts for 40% of the PAi calculation, was strongly phylogenetically and inversely distributed between phylogroup B1 and phylogroups A and B2 (**Figure 4.11**). However, it was surprising to observe that the combination with other phenotypes maintained this phylogenetic distribution. Additionally, strains from the newly described phylogroup F notably all had a relatively low PAi (**Figure 4.22B**). The examination of the distributions of the PAi according to collection and phylogroups confirmed these observations (**Figure 4.23**).

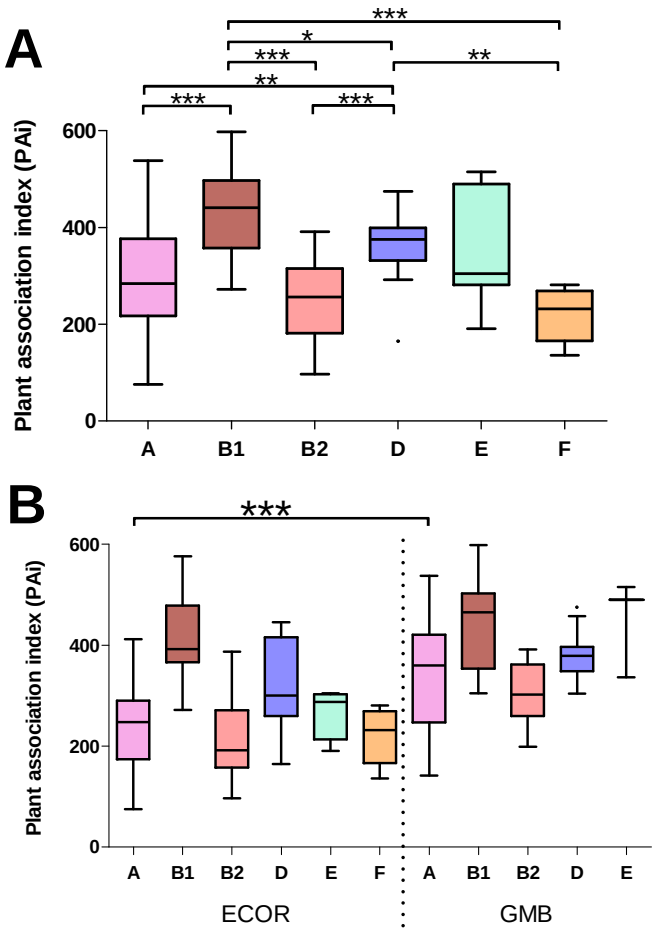


Figure 4.23. Distribution of PAi according to (A) phylogroups and (B) collection and phylogroups. Asterisks indicate significantly different means based on a Bonferroni post-hoc comparison after a one-way ANOVA ($p<0.0001$). The number of asterisks reflects significance threshold with *: $p<0.05$; **: $p<0.001$ and ***: $p<0.0001$. Significances between phylogroups were very similar between representations in panels A and B. For clarity, only the ECOR vs. GMB statistical significances were represented on panel B.

Similarly to what we showed for individual phenotypical rankings, we mapped PAi indices values onto the ClonalFrame phylogenetic tree of ECOR and GMB strains (**Figure 4.24**).

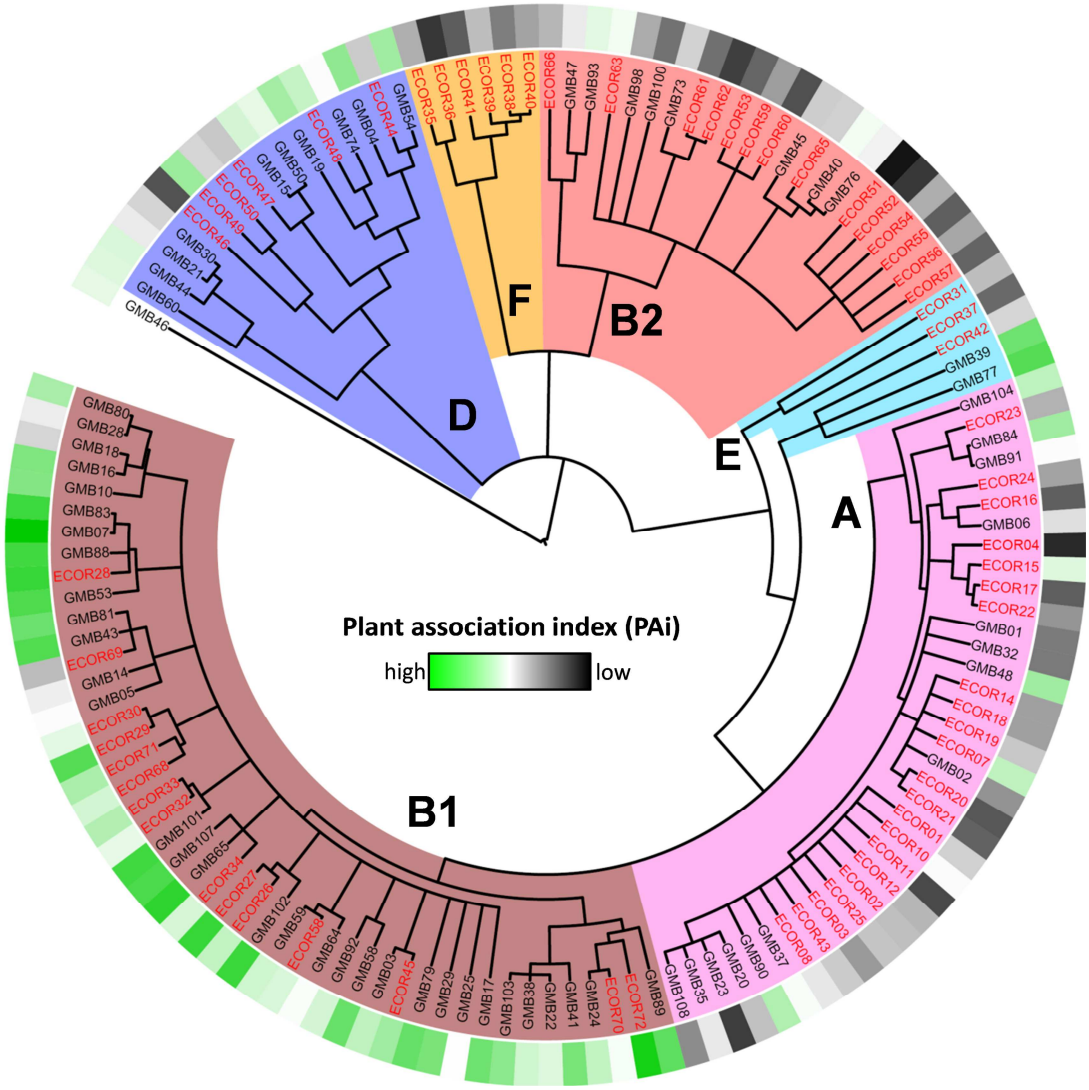


Figure 4.24. Distribution of the plant association index (PAi) on the ClonalFrame phylogenetic tree of ECOR and GMB isolates. As indicated in the middle of the tree, PAi is represented by shades of the following colours: green = highest PAi value and black = lowest.

Examination of the tree confirms visually that phylogroups B1, but also to some extent D and E are potentially more suited for plant association according to our selected criteria. On the other hand, phylogroup A, B2 and F strains had low levels of PAi. It is difficult to extrapolate on the direction of this adaptation: are B2 and A isolates evolving towards host specialisation, or are B1 isolates distinguishing themselves from host adapted phylogroups? An initial answer may result from the

phylogenetic reconstruction we performed and the overlap with PAi values, as presented in **Figure 4.24**. The most ancestral group is phylogroup D, from which phylogroups B2 and F, and then E, A and B1 diverged (**Figure 4.24**). If this phylogeny is correct, this would mean that D, having relatively high PAi values (**Figure 4.23**), is the most ancestral phylogroup, from which the others diverged. B2 and A would then have lost their high PAi, while B1 conserved it. However, if the phylogeny is as reported by Lecointre et al. (1998), placing B2 as the most ancestral group, it would then be D and B1 that evolved towards more host generalism and a higher nonhost adaptation.

It is nevertheless remarkable that an ecologically-relevant definition of possible environmental adaptation traits follow such a clear phylogenetical distribution. Our results taken altogether seem to strengthen the hypothesis that *E. coli* strains of various phylogenetic background and ancestries have different ecological behaviour. In the light of our observations, it is plausible that these differences are linked to different life histories in different ecological niches.

4.4.3. Ecological ranking of the *E. coli* O104:H4 2011 German outbreak strain

In May to June 2011, an outbreak of *E. coli* O104:H4 linked to the consumption of uncooked vegetable sprouts in Germany caused 50 deaths, 908 cases of haemolytic uremic syndrome and 3167 non-HUS cases of infection (Rohde et al. 2011). This outbreak was the largest and deadliest *E. coli* outbreak ever recorded. The genome sequences were quickly made public and a large crowdsourcing effort produced

valuable and rapid insights into the nature of this pathogenic strain (Rohde et al. 2011). As mentioned earlier in this thesis, we obtained the genome sequences of 2 related outbreak strains and, using MLST sequences, placed the outbreak strains in phylogroup B1. In the previous section, we found B1 strains to have singularities in terms of possible plant adaptive traits, and as this outbreak originated from fresh vegetables, it became very interesting to try and see where this strain would place in our ecological ranking of strains according to PAi. In August 2011, we received O104:H4 strain H1-1218 from the Health Protection Agency (HPA) and obtained its Biolog GN2 profile (**Figure 4.25**) thanks to Stephanie Schüller (IFR) who helped in manipulating the strain in a CL-3 laboratory.

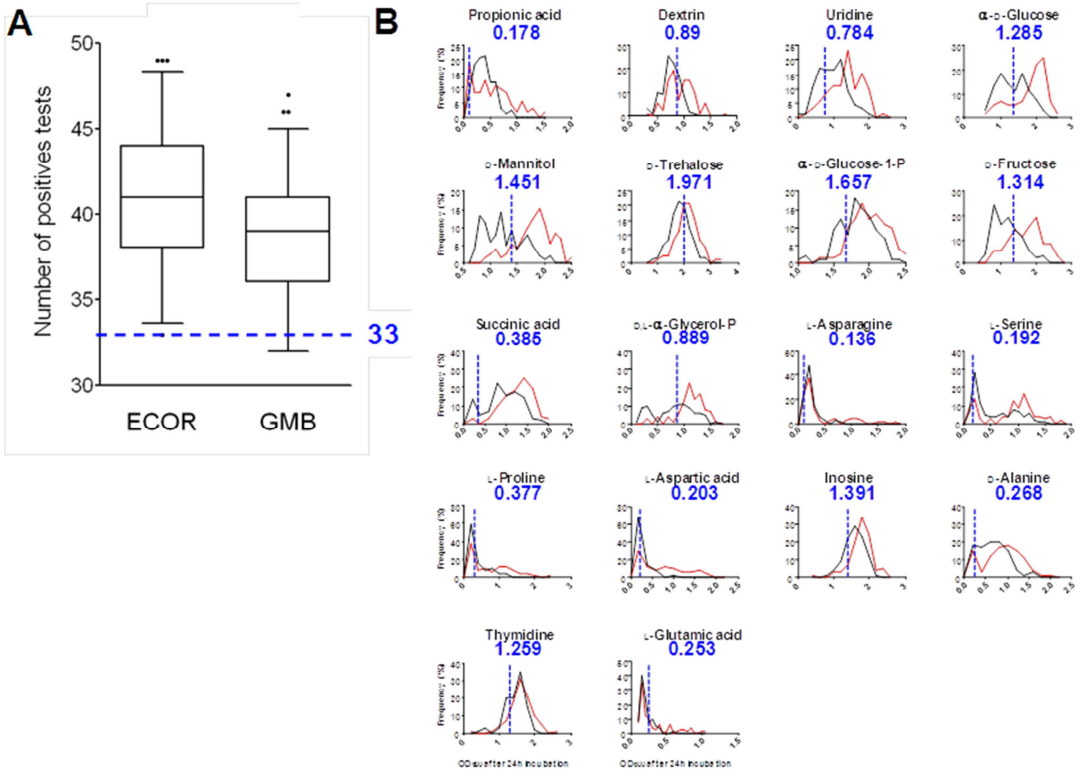


Figure 4.25. Characteristics and growth values on Biolog GN2 plates after 24h incubation at 37°C for selected C-sources by *E. coli* O104:H4 strain H1-1218 involved in an outbreak in Germany in 2011. Panel A represents Figure 4.4 and panel B represents Figure 4.7 presented in this thesis. The values for strain H1-1218 are indicated in blue.

Strain H1-1218 was overall a very low metaboliser, able to reach OD_{600}>0.63 after 24 h of incubation at 37°C on 33 C-sources only, which places it at a lower level than any ECOR strain, and among the bottom 25% of all GMB strains (**Figure 4.25A**). When C-sources shown to discriminate ECOR and GMB metabolic profiles to the greatest extent were examined, H1-1218 was always amongst the lower metabolisers (**Figure 4.25B**), which was also observed when looking at sucrose and *p*-HPA metabolism (**Figure 4.26**).

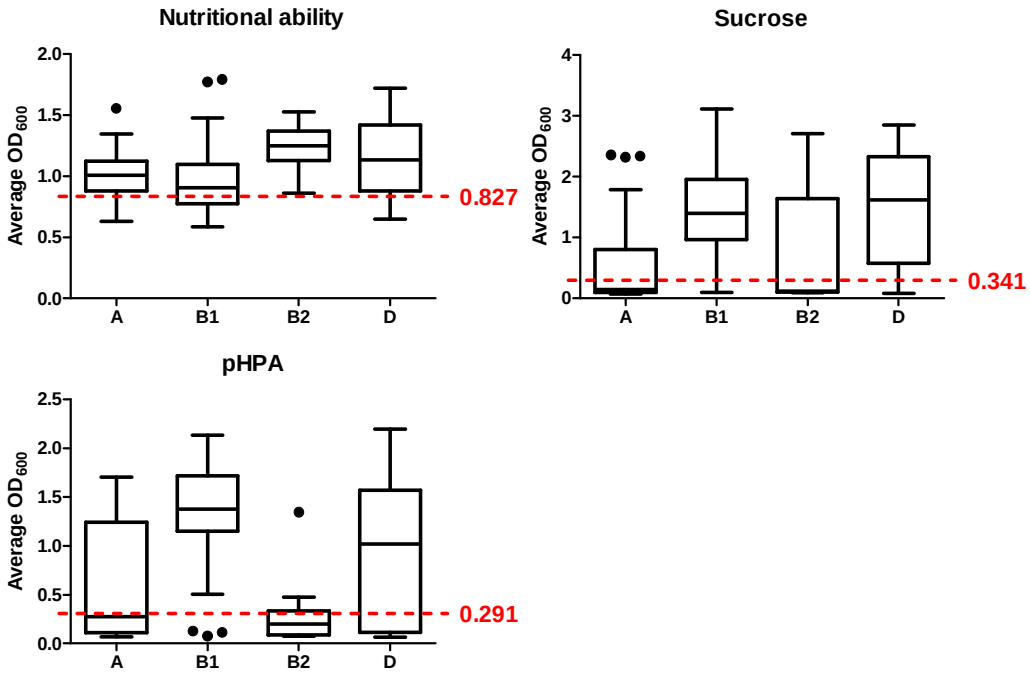


Figure 4.26. Nutritional abilities, sucrose and *p*-HPA metabolism as used to define the PAi and comparison of values of *E. coli* O104:H4 strain H1-1218 involved in an outbreak in Germany in 2011 after 24h of incubation on Biolog GN2 plates at 37°C. H1-1218 values are indicated in red. "Nutritional abilities" represents the average of Figure 4.25B.

Surprisingly, we observed no growth after 24 h on sucrose or on D-serine or *p*-HPA. After examining the public genome sequence of H1-1218, we could identify the presence of sucrose metabolic *csc* genes disrupting the D-serine utilisation genes, but

we could not find the *hpa* cluster for *p*-HPA utilisation. However, when incubated for more than 24 h on Biolog GN2 plates, H1-1218 could indeed grow on sucrose (reaching OD_{600}=1.697 at 72 h), but not D-serine or *p*-HPA, in accordance with genomic observations (data not shown). Interestingly, the growth of H1-1218 on sucrose was among the slowest we observed (data not shown), suggesting possible differences in the metabolic utilisation of sucrose between the O104:H4 strain and the rest of our collection.

Unfortunately, because of time and safety constraints, we were not able to quantify biofilm formation of the O104:H4 pathogenic strain, so we could not calculate its PAi as we did before. However, we can account for 60% of its PAi (the metabolic factors; **Figure 4.26**): nutritional abilities would place the O104:H4 strain at rank 35/170, sucrose metabolism at 110/170 and *p*-HPA metabolism at 112/170, which means that even if the O104:H4 strain had the top ranks for biofilm formation and siderophore production, it would be ranked 22/170 for PAi.

From this seemingly poor "plant association index" according to our criteria, we can speculate that the O104:H4 strain may not be very well adapted to nonhost environments, or did not persist long in the environment before living in or infecting its hosts. However, it is also plausible that the observation that the overall metabolic abilities are very low at 24 h (**Figure 4.25**) is a direct consequence of the stress resistance abilities being very high, as a result of the SPANC balance introduced earlier (section 4.2.4 of this thesis). These hypotheses have to be verified experimentally, but in any case the fact that this outbreak strain does not perfectly

"fit" the expected profile we assembled from our collection of nonhost strains from plants is informative.

4.5. Conclusions and industrial relevance

4.5.1. Ecological considerations and possible link between genetic regulation and genome dynamics

In summary, we were able to observe clear phenotypic differences between plant (GMB) and faecal (ECOR) isolates. We first observed variation in metabolic abilities, as monitored using Biolog GN2 plates. The biggest difference we observed was that ECOR strains were faster metabolisers than GMB strains, especially when growing on amino acids as sole C-source, and could reach much higher OD_{600} values on average after 24 h incubation on common C-sources such as glucose, fructose or mannitol.

As mentioned in previous sections, it is possible that this observed lower metabolism is an indirect consequence of a higher stress resistance by GMB strains (Ferenci 2005), which we unfortunately did not test for in this work. This speculation is ecologically plausible as nonhost environments, even if estimated to harbour half of all living *E. coli* (Savageau 1983), are presumably more stressful environments than the gastrointestinal tract of mammals, the primary reservoir of *E. coli*. Additionally, GMB strains were in proportion more able to use sucrose and the aromatic compound *p*-HPA and this ability was strongly correlated with phylogroups B1 and D. These two C-sources have tremendous ecological significance. Aromatic compounds are believed to be most abundant and available to *E. coli* in soils (Diaz et al. 2001) and

sucrose is the most abundant carbohydrate in plants and vegetation. The primary environment of *E. coli* being the mammalian gastrointestinal tract, conceivably soil, vegetation and water are the most common nonhost environments encountered by any excreted *E. coli*.

Looking at a range of colonisation-associated phenotypes, we also observed that GMB strains were on average more motile, could produce more biofilm *in vitro* but were poorer siderophore producers than ECOR strains. These observations that flagellar motility, sucrose metabolism are potentially important factors for *E. coli* life on plants and that amino acids metabolism is affected bore interesting similarities with the transcriptional profile of *E. coli* O157:H7 strain EDL933 grown in the presence of shredded lettuce juice (Kyle et al., 2010). In this work, the most highly upregulated categories of orthologous genes (COG) in the presence of plant lysates were cell motility and the related intracellular trafficking and secretion, and sucrose metabolism genes (Kyle, Parker et al. 2010; Maria T. Brandl, unpublished observations). Surprisingly, amino acids and nucleic acid transport and metabolism genes were globally and most severely downregulated. The combination of our results with the study on the transcriptomic response to lettuce lysates (Kyle, Parker et al. 2010) also stresses that important traits required by *E. coli* pathogens to colonize plants may not necessarily be pathogenesis-associated functions as sometimes suggested (Berger et al. 2010), but functions that are commonly shared within the *E. coli* species, possibly from the flexible gene pool. This also suggests that there is a link between the transcriptomic response to environmental conditions and the shaping of genome contents in populations associated with the corresponding environment. Environments that provide advantages to a certain metabolic ability cause an

upregulation of the corresponding metabolic genes, providing the strain possess them as selective advantage. In that sense, it sounds plausible that gene regulation has a role to play in genome dynamics, the shaping of gene content in bacterial genomes and possibly the different types of selection pressures that can be observed on various parts of the genome.

4.5.2. Industrial relevance

The most significant meaning of the PAi is that it represents a combination of traits characterising *E. coli* isolates according to their fitness in secondary environment, with B2 and A having the lowest values and B1 and to some extent D having the highest (**Figure 4.23**, **Figure 4.24**). As presented in the Introduction (section 1.1.3.2), strains from phylogroup B2 and A seem to be evolving toward an increased host specialisation, whereas strains from phylogroup B1 seem to be generalists, without any clear host preference (White et al. 2011). Previously published studies suggest that *E. coli* phylogroups differ in their host association abilities, as some phylogroups are host specialists (the archetype being B2) whereas others seem to be host generalists (B1) (White et al. 2011; Sims and Kim, 2011). The analyses presented in this chapter contribute to further this dichotomy, by showing that host generalisation in phylogroup B1 is additionally associated with nonhost adaptation. In a study involving intestinal colonisation of infants, phylogroup B2 has been shown to be much more frequent among resident strains, persisting for long periods in hosts as opposed to transients, only transiting for a few days or weeks in the intestines (Nowrouzian et al. 2005). In this study, B1 was in very low quantities in residents,

and slightly more in transient, but reflected an overall poor host association (Nowrouzian et al. 2005).

Taken altogether, our observation and the presented literature can highlight the very interesting possibility that PAi levels can predict both plant and host adaptedness, and thus the possibility for any given strain to be more likely to be either a resident or a transient if ever recolonizing a host. For the food industry, this has a potentially important impact, as *E. coli* from food are potentially ingested. When currently monitoring for *E. coli*, the industry is aware of the levels of *E. coli* on its food products, and is trying to keep it as low as possible. Conceivably, further information could be obtained, as it can be imagined that the population structure of isolates from specific locations (e.g., agricultural fields) is monitored, and would provide additional useful data on (a) the overall likelihood of successful recolonisation if the produce is eaten as well as (b) the possible source of contamination (see Chapter 3 section 3.6). Of course, this hypothesis relies on the combination of published data by other groups as well as interpretation of our results, and it remains important to confirm it experimentally.

5. Dynamics of *E. coli* colonisation of salads and interactions with the natural phyllosphere microflora

5.1. Context of this study

The agricultural field environment is heavily populated by soil-borne and plant-associated microbes. Being outdoor open environments, daily temperature fluctuations, but also variation in chemical contents of soils, humidity or radiation exposure are common and impact on the structure of natural microflorae and functional communities. When immigrating bacteria colonise agricultural fields, they have to interact with an existing community presumably fully adapted to versatile and stressful conditions, with which competition for resources has to be fierce. However, colonisation by invading bacteria, including plant or human pathogens, can be successful prompting questions about their interactions with the resident microflora, and the mechanisms they developed to maintain an environmental fitness conducive to growth and survival. It has been suggested that epiphytic bacteria could stimulate or suppress colonisation by plant pathogenic bacteria (Lindow and Brandl, 2003). Conceivably, epiphytic bacteria could have a similar role on the persistence of human pathogens and commensals colonising plants.

In order to gain insight on how colonising human pathogens and gut-adapted bacteria such as *E. coli* can persist and successfully colonise plants, it is relevant to investigate their dynamics of nonhost contamination, as well as their interactions with the rhizosphere or phyllosphere natural microbial communities. *E. coli* O157:H7 has been shown to grow in vegetables (Li et al., 2001; Cooley et al., 2003) or at least persist for

long periods (Solomon et al., 2003; Islam et al., 2004; Ibekwe et al., 2007). Additionally, studies focused on the microbial ecology of plants have reported a significant reduction of *E. coli* counts in the presence of epiphytic bacteria (Cooley et al., 2006) or even alteration of the microbial community structures caused by *E. coli* (Lopez-Velasco et al., 2010). However, there is limited knowledge on this last point, even if preliminary studies seem to indicate that there could be an explicative link between how *E. coli* can colonise certain salad crops and how natural phyllosphere communities react to such a colonisation (Lopez-Velasco, Davis et al. 2010).

In this chapter, we investigate that link by experimentally contaminating spinach grown either in the field or in controlled laboratory conditions. Here, we present analyses of the salad colonisation dynamics by *E. coli* and of the interactions between the natural microbial communities living in agricultural soils and field-grown plants and contaminating *E. coli*.

5.2. Testing of DGGE protocol using field samples

In this chapter, we present data resulting from experimental contamination with *E. coli* of field-grown plants. As our access to field-grown plants was limited to summer, we could not spend a lot of method development time on the actual experimental contamination and decided to test the DGGE protocol on soil and leaf samples from regular commercial spinach and rocket salad grown in fields in Norfolk (**Figure 2.14**).

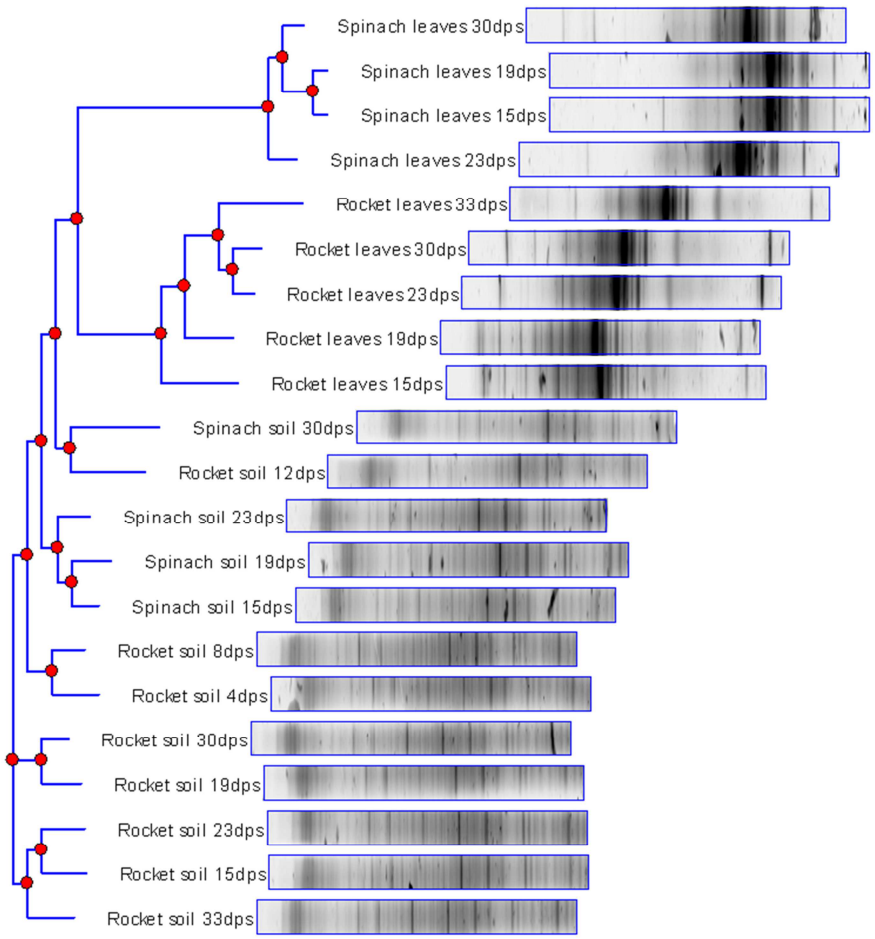


Figure 5.1. Relatedness between representative community profiles of field-grown spinach and rocket salad. This dendrogram was calculated with the neighbour-joining method on the Pearson correlation between individual profiles of the same 16S rDNA DGGE gel. “dps”: days post-sowing. This figure was created using Nonlinear Dynamics Phoretix 1D v11.1 (demo version).

The protocol for DNA extraction from soil and leaf material and the subsequent PCR-DGGE worked well as we observed multiple bands in each profile from soil or leaves (**Figure 5.1**). We observed a higher richness (number of bands) in soils than on leaves (Mann-Whitney U=2.00; p=0.0003). Rocket and spinach soil profiles showed averages of 40.75 and 37.5 distinct bands respectively, compared to 23 and 14.5 bands in rocket and spinach leaves profiles. This difference in richness can of course be caused by different DNA extraction protocols, but it also possibly highlights that soils generally harbour higher levels of microbial diversity, as they offer a more protected environment than aerial parts of plants, which are more exposed to outdoors variation and subject to more intense stress exposure. Consistently, we observed a very clear grouping according to the origin of the profiles (**Figure 5.1**). Samples from rocket and spinach leaves clustered separately, and both leaves-associated profiles were more distant than from soil samples, which also tend to weakly correlate to the type of plant grown in them (**Figure 5.1**).

This simple test of protocol indicated that we had chosen a suitable way to monitor the dynamics of bacterial populations on leaves (and soil), which could be applied to our experimental contamination investigation plan.

5.3. Experimental *E. coli* contamination of spinach plants grown in controlled conditions

5.3.1. Spinach cultivar used in this study

We asked our partner farms for different spinach cultivars to grow in our laboratory-scale experiment and they provided us with seeds from cultivars Toucan, Lazio, Picasso, Douglas and Sardinia. To determine which of these varieties would be the easiest to handle in laboratory conditions, we tested their germination abilities at 20°C (**Figure 5.2**).

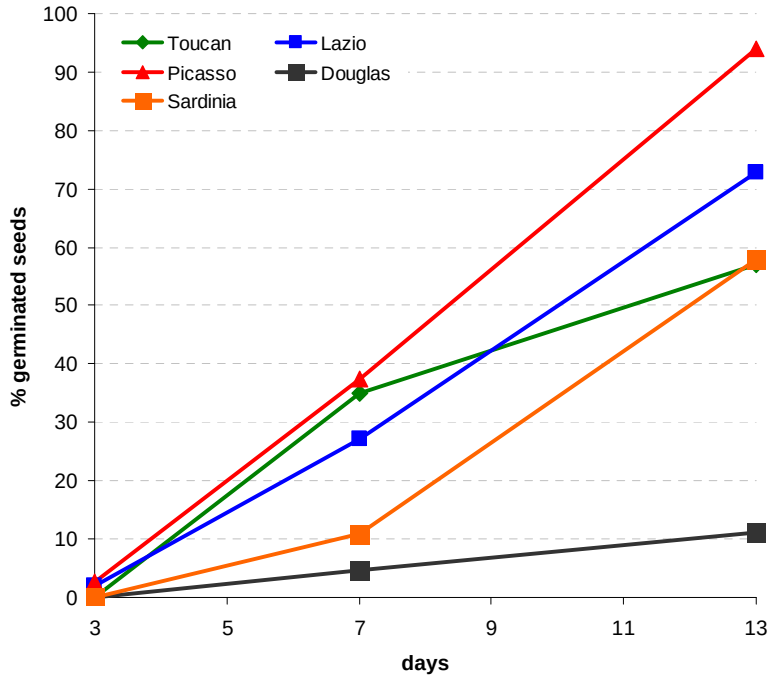


Figure 5.2. Germination frequencies of different spinach cultivars. Seeds (n=~100) from 5 cultivars were placed on filter paper in sterile Petri dishes and watered regularly for 13 days. The number of germinated seeds was determined after 3, 7 and 13 days.

Cultivar Picasso had the higher germination rate at 3 days and 13 days (**Figure 5.2**), and was the cultivar we chose for culture chamber growth. We collaborated with staff and facilities from the John Innes Centre Horticultural Centre (Norwich, UK) to grow spinach in culture chambers with controlled temperature, humidity and light cycles. Plants were grown in non-sterile soil to maximise the likelihood of getting a representative and biologically meaningful microflora on leaves. It was confirmed by the JIC Horticultural Centre staff that cultivar Picasso was the fastest to germinate in soil, although cultivars that were observed not able to germinate well *in vitro* (i.e. Douglas) could also, to some extent, germinate better in soil (data not shown).

As an additional control step and to determine if *E. coli* could be present in or within commercially available seeds, we crushed roughly 250 seeds from each cultivar, placed the powder into BGLBB medium for *E. coli* enrichment and plated after incubation on TBX medium. We did not retrieve any blue colony on TBX (data not shown) indicating that *E. coli* was not likely to be present in any of our seed batches and would not perturb our colonisation experiments.

5.3.2. *E. coli* strain used in this experiment

As this study was scheduled to be performed before performing most of the experiments and analyses presented earlier in this thesis, we had limited knowledge on the GMB strains to select one with relevant properties (for instance, metabolic). We chose strain GMB30 as the main strain used in our experimental contamination experiments, based on the decision that we would like to use a strain arguably representative of the "average plant-associated isolate". We then based our selection

at the time on the fact that GMB30 scored averagely in biofilm experiments (rank 50/173). Retrospectively and after the DGGE experiments were performed, we could examine how GMB30 performed in all other assays: siderophore production (rank 101/173), nutritional abilities (109/173), sucrose utilisation (76/173) and *p*-HPA utilisation (149/173). Combined together, these values indicated a PAi rank of 118/173, placing the strain slightly above the bottom 30% of strains that are presumably less adapted for plant persistence, according to the criteria described in Chapter 4.

At first, this retrospectively indicated to us that GMB30 was perhaps not the best choice for our experiment. Nevertheless, we observed interesting results with this strain, as described in this section.

5.3.3. Colonisation dynamics of *E. coli* on laboratory-grown spinach

From most of the studies looking at interaction between *E. coli* and plants, it is assumed that *E. coli* can infect plants from an original epiphytic contamination (irrigation, rain splashes, or faecal contamination of avian origin). We also followed this assumption in our experiment by inoculating *E. coli* GMB30 by spraying leaves of grown plants. We then sought to know if GMB30 was actively growing, surviving or dying after being sprayed on spinach plants. During 12 days, we monitored bacterial counts on 3 infected plants per time point by plating bacterial suspensions on violet red bile with glucose (VRBG) agar medium, selective for bile-resistant coliforms (**Figure 5.3**). Negative control plates on uninfected plants did not show any

E. coli count on VRBG, although we could isolate colony morphotypes distinct from *E. coli* throughout the experiment (data not shown).

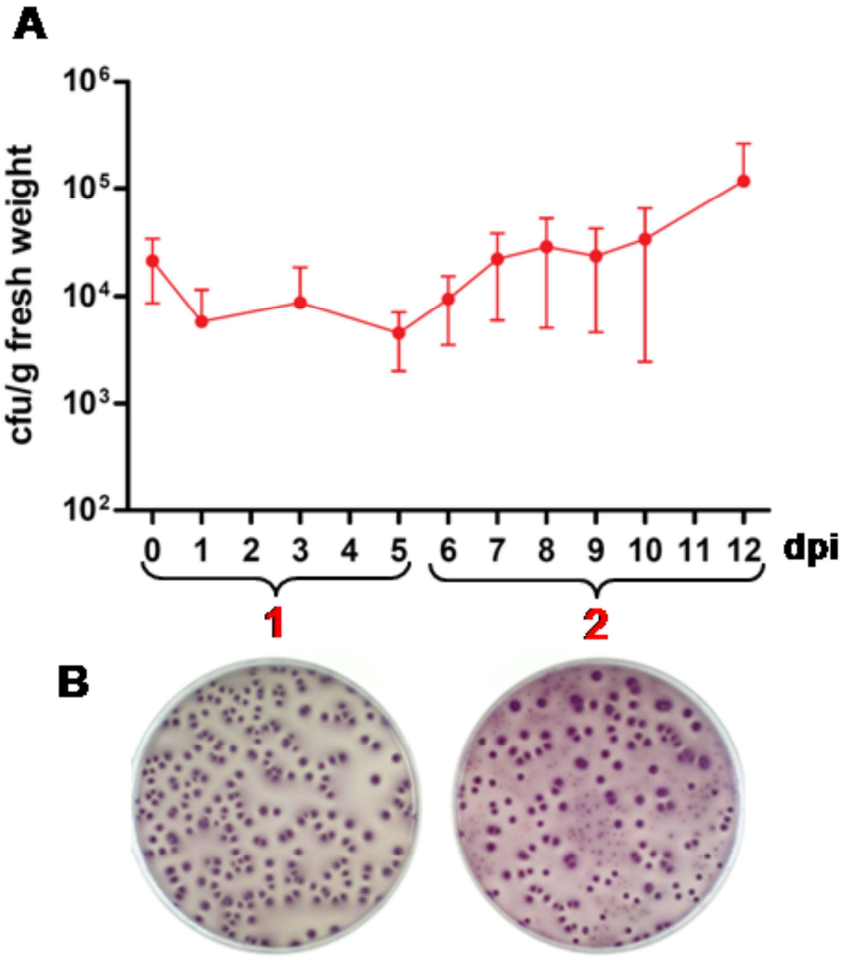


Figure 5.3. Colonisation dynamics of *E. coli* GMB30 infecting spinach cultivar Picasso plants grown in controlled environmental conditions. (A) Bacterial counts from 0 to 12 days post-infection (dpi). Error bars represent the standard deviation from the mean after 3 technical replicates. (B) Pictures of representative VRBG plates obtained during counts presented in panel A. "1" and "2" represent the two distinguishable phases in which we obtained different types of colonies growing on agar plates.

We observed that GMB30 counts were first declining and stabilising (days 0 to 5 after infection) and then increasing (days 6 to 12 after infection). However, to our great

surprise, we could reproducibly get two different kinds of colony patterns on VRBG plates, depending on the days after infection by *E. coli* (**Figure 5.3B**). From days 0 to 5 after infection, we observed purple *E. coli*-looking colonies growing homogenously on VRBG but from days 6 to 12 after infection, we observed the same colonies in majority, but mixed with different other colonies of different sizes, colour and smell. This observation was first dismissed as a possible contamination during our processing of the plants on the grounds that we reproducibly observed it specifically occurring roughly a week after *E. coli* infection. Additionally, we never observed such high counts of non-*E. coli* strains on negative control plates, from plants sprayed with only water (data not shown) suggesting that this could be an *E. coli*-specific effect. When identified by 16S rDNA sequencing, these colonies were identified to be mostly from genus *Stenotrophomonas sp*. (family: *Xanthomonadaceae*) and genus *Rahnella sp*. (family: *Enterobacteriaceae*). These genera are usually associated with soil or water environments, but are also found in great quantities on plants in agricultural settings (Suckstorff and Berg, 2003; Ryan et al., 2009; Vyas et al., 2010). Additionally, *Rahnella sp.* are usually psychrotrophic bacteria that are often involved in produce spoilage (Randazzo et al., 2009).

This observation suggests that contamination by *E. coli* can perturb the structure of bacterial populations residing on spinach plants. Because we saw counts of bacteria such as *Stenotrophomonas* and *Rahnella* increasing, this also suggests that *E. coli* was somehow able to enhance the growth of these bacteria, maybe by synthesising a growth-promoting substance.

5.3.4. Perturbation by colonising *E. coli* of natural resident communities associated with spinach

We reproduced these resident bacterial population perturbations observed during spinach experimental *E. coli* contamination, this time by extracting total DNA from the phyllosphere and monitoring bacterial community dynamics using DGGE. The main objective of this study being to understand how *E. coli* interacts with natural phyllosphere microbial communities in the agricultural field environment, we performed spraying experiments on both field-grown plants (n=2), and plants grown in controlled environmental conditions (n=5), monitoring bacterial communities from 1 dpi to 8 dpi by DGGE. In this section, we will show the results of one spraying experiment on field-grown plants and 2 spraying experiments on laboratory-grown plants.

5.3.4.1. Experimental contamination of field-grown spinach

We transferred 3-weeks old field-grown spinach plants into individual pots and brought them back in the laboratory for spraying with *E. coli* GMB30. We sprayed 5mL/plant of bacterial suspensions containing different concentrations (10^4 to 10^7 cells per mL) of *E. coli* GMB30. After 1, 4 and 7 days, we extracted bacterial DNA from leaves and amplified the V3 region of the 16S rDNA gene to use for DGGE (see Methods). We obtained the gel shown in **Figure 5.4**.

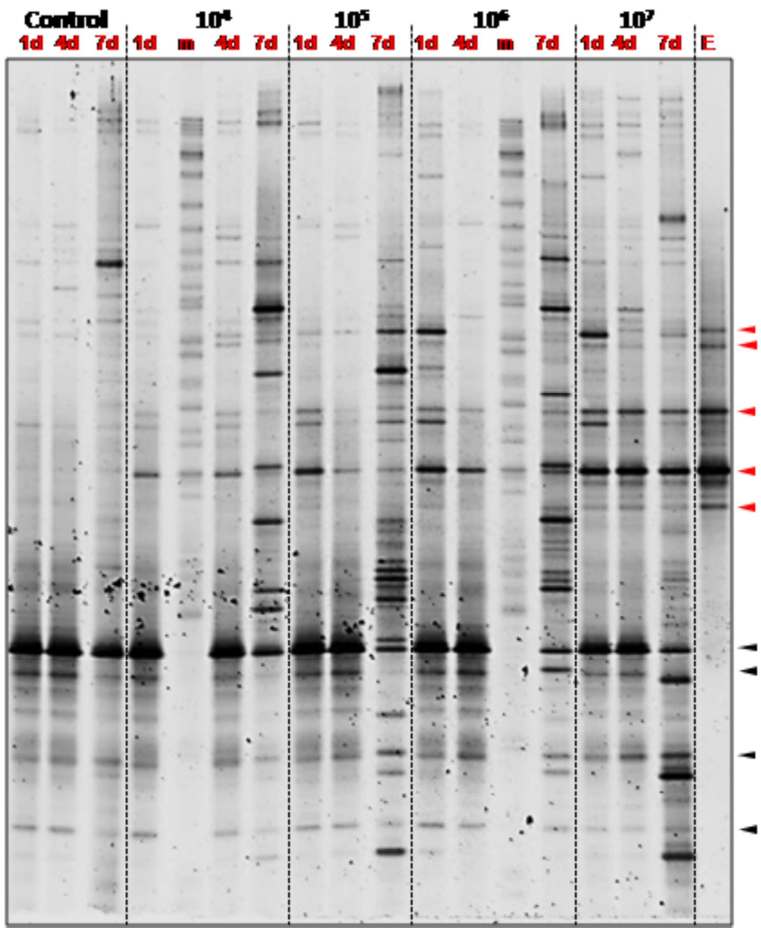


Figure 5.4. Denaturing gradient gel electrophoregram showing community-profiling of field-grown spinach sprayed with different concentrations of *E. coli* strain GMB30. Red arrows indicate the major bands amplified from a single pure culture of *E. coli* GMB30 (Lane "E"); black arrows indicate the major bands of chloroplastic DNA; "d": days post-infection (with *E. coli*); m: clone ladder prepared and given by P. Tourlomousis and used solely for gel analysis.

First, we observed that the profile from *E. coli* GMB30 pure culture was not composed of a single band, as we expected. The fact that a V3 amplification of pure cultures of *E. coli* yields multiple bands has been reported in another study, in which a thorough analysis is presented (de Araujo and Schneider, 2008). It has been suggested that the observation of multiple band amplification from pure cultures is caused by the fact that *E. coli* has 7 copies of ribosomal genes having different sequences (Kang et al., 2010) and being amplified in the same reaction step (de Araujo and Schneider

2008). Because of this phenomenon (i.e. different bands not necessarily representing different bacterial species) DGGE is probably not a very accurate method for inferring detailed bacterial abundance (de Araujo and Schneider 2008) although, it is a very flexible and quick method to compare community profiles accurately and efficiently (the variation in abundance of one band across comparable profiles, regardless of its biological significance, is interesting). In our case, a major band was present and possibly representing the best amplification, or amplification of similar sequences in different copies. Nevertheless, we observed that these bands were expectedly present in all infected samples as early as 1 dpi and not in the controls (**Figure 5.4**).

In confirmation with our expectations based on plate counts, we observed very notable increases in the number of bands at 7 dpi compared to 4 dpi or 1 dpi in the *E. coli*-treated samples, but not in the controls (**Figure 5.4**, **Figure 5.5**). This increase in band number was observed regardless of inoculum concentration. Interestingly, although the bands appearing at 7 dpi were at the same position for each tested inoculum concentrations, their intensity seem to vary very much (this point is particularly well illustrated when comparing profiles at 7 dpi between inoculums 10^5 and 10^6 on **Figure 5.4**).

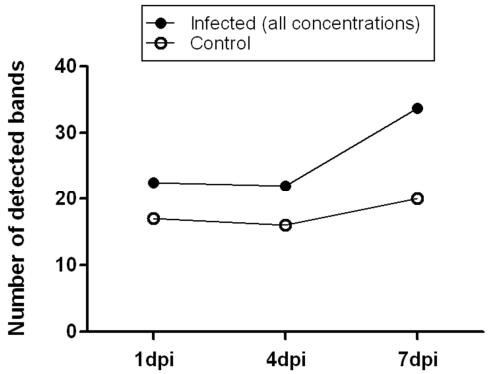


Figure 5.5. Number of detected bands for plants infected vs. non-infected with *E. coli*. Infected samples represent the average of all inoculum concentrations; "dpi": days post-infection.

To see how community profiles were related to each other, we grouped similar samples according to the Pearson correlation of their DGGE profiles in a neighbour-joining dendrogram (**Figure 5.6**).

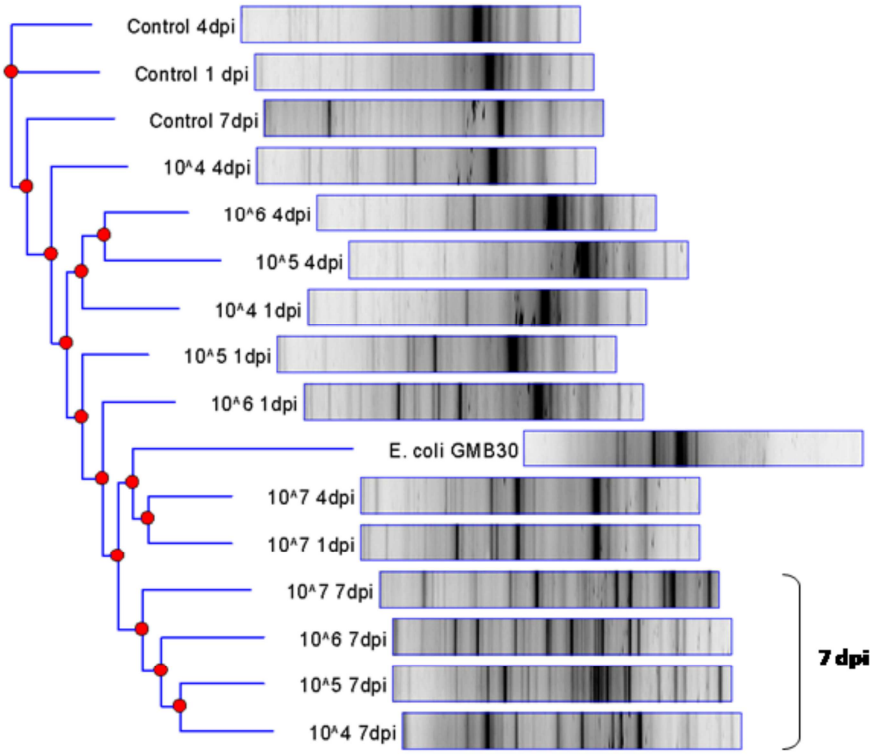


Figure 5.6. Relatedness between community profiles of field-grown spinach sprayed with different concentrations of *E. coli* strain GMB30. This dendrogram was calculated with the neighbour-joining method on the Pearson correlation between individual profiles of the same 16S rDNA DGGE gel. "dpi": days post-infection (with *E. coli*). This figure was created using Nonlinear Dynamics Phoretix 1D v11.1 (demo version).

We expectedly observed that infected 7 dpi sample profiles clustered together and were the most different profiles from the rest (**Figure 5.6**). Additionally, the control sample profiles also grouped together (**Figure 5.6**) except for the control at 7 dpi, which although similar, showed a slight increase in its number of bands. Profiles of infected plant samples at 1 dpi and 4 dpi generally grouped together, with the notable

exception of the 1 dpi and 4 dpi samples infected with 10^7 cells per mL, which grouped with the 7 dpi samples (**Figure 5.6**). This last observation is very interesting, as it suggests that if the increase in bands is indeed caused by the addition the *E. coli*, it is also faster when more *E. coli* are added. In other words, the higher the inoculum concentration is, the more perturbed are profiles at 1 dpi and 4 dpi compared to uninfected profiles.

Unfortunately for this experiment, we did not perform plate counts of *E. coli* colonising plants. This information could have helped us to determine if the differences in profiles a week after infection was linked to *E. coli* growth patterns on leaves, or was even similar to what we observed in **Figure 5.2**.

5.3.4.2. Experimental contamination of spinach grown in controlled environmental conditions

We attempted to perform another experimental contamination using plants grown in another field, but half-way through the experiment we observed heavy fungal or oomycete contamination on our spinach plants, which we discarded. We therefore decided to continue growing spinach in laboratory conditions (as we had initially observed the microflora perturbation) to further investigate this phenomenon.

In **Figure 5.7**, we present profiles at 4 to 8 dpi for 2 representative experiments (labelled “S2” and “S3”), to which we added profiles from the inoculated field-grown plants (labelled “S1”) presented in **Figure 5.4** and **Figure 5.6**.

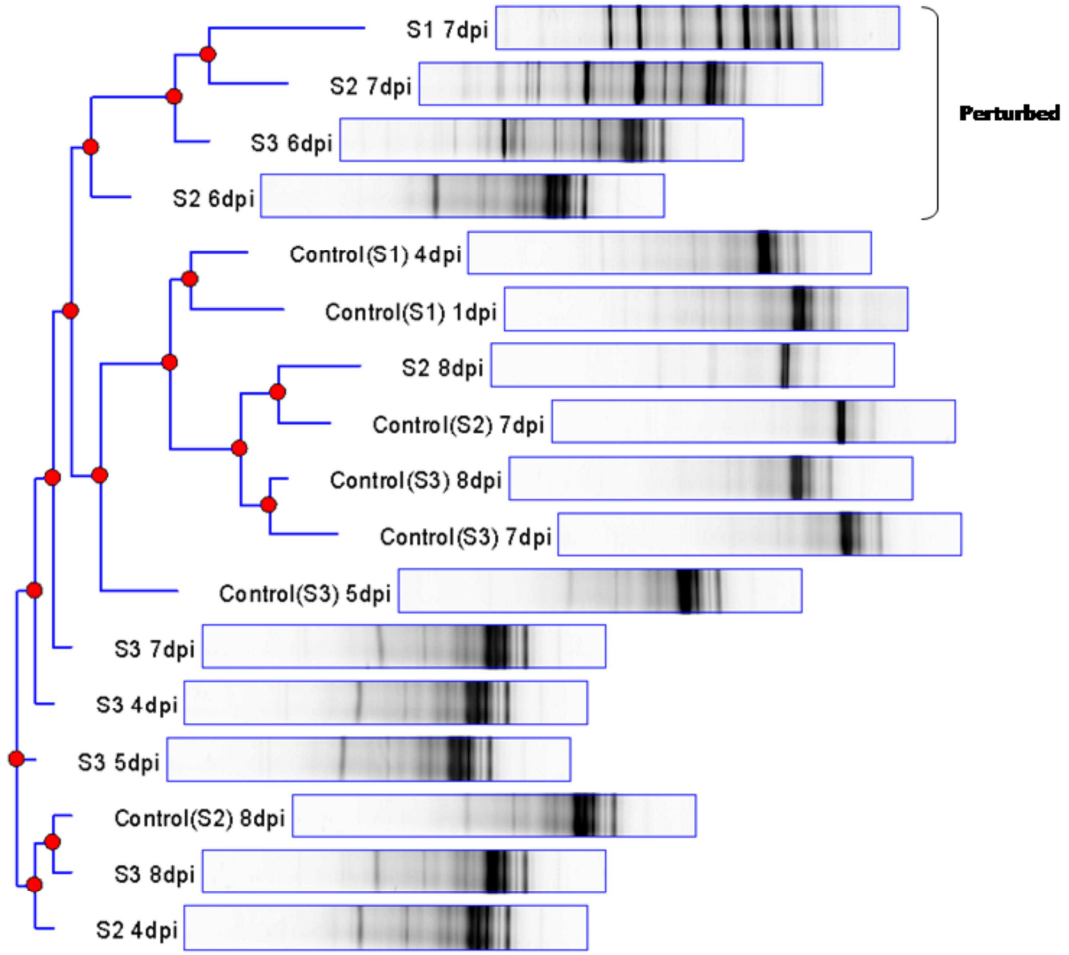


Figure 5.7. Community-profiling of field-grown (S1) and culture-chamber-grown (S2 and S3) spinach sprayed with *E. coli* strain GMB30. This dendrogram was calculated with the neighbour-joining method on the Pearson correlation between individual profiles of the same 16S rDNA DGGE gel. "dpi": days post-infection (with *E. coli*); "S" stands for "spraying experiment". This figure was created using Nonlinear Dynamics Phoretix 1D v11.1 (demo version).

Similarly to field-grown plants, we observed an increase of the number of bands (and thus possibly bacterial diversity) about a week after inoculation of culture chamber-grown plants. Profiles from the 3 spraying experiments showing an increase in number of bands were the most correlated and clustered together in the dendrogram (**Figure 5.7**), despite the fact that the appearing bands were not similar between field-

and culture chamber-grown plants. Interestingly, the difference occurred one day earlier (6 dpi instead of 7 dpi), which is possibly due to bacteria being able to grow faster or at least survive better in a propagator with a constantly controlled temperature (22°C) rather than daily fluctuating temperatures in field conditions. Also, there did not seem to be an observable change in community structure from 4 to 7 dpi. Profiles from non-inoculated plants also expectedly grouped together, despite their significant difference in number of bands (**Figure 5.8**). Indeed, non-inoculated plants grown in the field had much more bands on average (15.8) than culture chamber-grown plants (6) and this difference was statistically significant (Mann-Whitney U=1.5; p=0.0278).

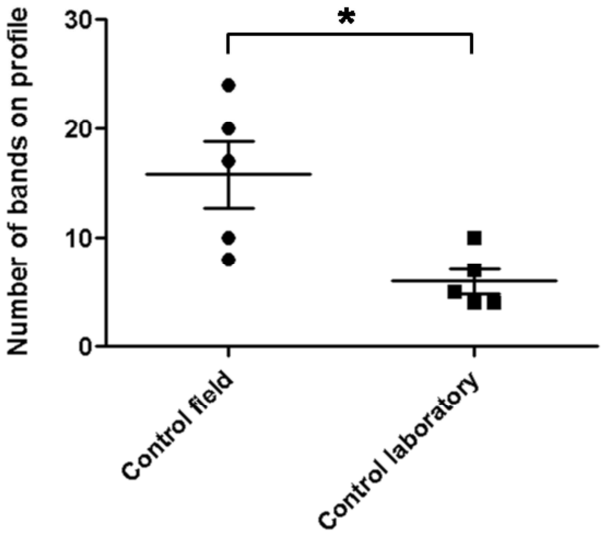


Figure 5.8. Comparison of community profiles between non-inoculated field-grown and laboratory-grown plants using DGGE. The number of bands were determined from experiments shown in the previous sections. Each point represent the number of bands of a profile from non-inoculated plants. The asterisk indicates a significant difference after a MWW test.

This observed difference in the number of bands is possibly caused by varying levels of bacterial richness in field and controlled environments. When growing in fields,

plants are subjected to irrigation from above, the sustained presence of wildlife and birds and heavily varying meteorological conditions. It is therefore not surprising that the bacterial richness in such dynamic environments is observed to be bigger than the bacterial richness on plants grown in very controlled conditions. The fact that we did not observe the same bands appearing in field- and laboratory-grown plants infected with *E. coli* is a good indication that the natural microflora, assumed to be different in different growing conditions, is perturbed. We do not see any perturbation in non-inoculated plants, which is also supporting this hypothesis.

5.3.4.3. Comment on the amplification of contaminating eukaryotic DNA

We very consistently observed very bright bands in the phyllosphere samples, around 54% denaturing conditions (created by urea and formamide in an 8% acrylamide gel) for spinach and around 50% denaturing conditions for rocket salad. These bands were identified as contaminants from chloroplastic and mitochondrial DNA by comparison with other studies (Lopez-Velasco et al., 2010; Rastogi et al., 2010) although mitochondrial contamination was estimated to be low for salad samples (Rastogi, Tech et al. 2010). Judging by some of our profiles, this contamination can represent up to around 80% of total amplicons in abundance, which can lead to an overestimation of microbial abundance in certain profiles. However, in our analysis, we focused on the position and the number of bands in each profiles rather than their weight. Indeed, DGGE has been considered as a semi-quantitative method for comparison across samples on different gels (Tourlomousis et al. 2010) but we preferred to stay cautious and not use densitometric information from the gels. 16S

rDNA PCR-based community analyses do not tend to reflect very accurately the abundance of different species, but more the efficiency of the annealing of universal primers to particular templates. Furthermore, it has been estimated that PCR misses half of the ribosomal diversity in environmental samples, because of annealing differences (Hong et al., 2009).

Nevertheless, we tried to investigate whether we could get rid of these contaminants. It has been recommended to use primer pairs targeting different hypervariable regions to overcome this unwanted amplification. We tried to compare amplification targeting the V1-2, V1-3, V3, V3-5, V6-8 and V8 hypervariable regions (see Methods) and although we did not obtain the same bands (but roughly the same number) we equally observed a majority of chloroplastic DNA in the amplicons for both hypervariable regions targeted, as indicated by wider bands on the denaturing gels (data not shown). Additionally, we did not find differences in amplification when the annealing temperature was raised from 50°C to 55°C. We then tried different methods for microbial DNA retrieval, with sonication or stomaching of leaf material (mechanical disruption of leaves). In both cases we did not get rid of chloroplastic material (data not shown). This observation has been confirmed later by a study where authors could amplify plant DNA after just gently washing the surface of leaves with diluent (Rastogi, Tech et al. 2010). A clever way to remove these contaminants has been suggested in the same recent study. They identified recognition sites of uncommon restriction enzymes in the lettuce chloroplastic 16S rDNA sequence. After extraction from environmental samples, the DNA was digested using these restriction enzymes and the band corresponding to mostly undigested bacterial 16S rDNA fragments was excised to be used for DGGE. The advantage of such a method is that providing

digestion is complete, contaminant chloroplastic DNA is theoretically completely removed. A disadvantage however is that one cannot assert that fragments of bacterial origin do not get digested as well (Rastogi, Tech et al. 2010).

In this study, because of the preliminary nature of the analysis and the lack of a high number of samples and field-scale replicates, we largely focus on the comparison of community profiles alone. Our assumption was that as long as contamination by eukaryotic DNA is more or less similar across samples, it should not bias too much our interpretation and the relevance of the results presented here.

5.4. Conclusive remarks and industrial relevance

5.4.1. Ecological hypotheses

A study including very similar results and denaturing gel pictures than those presented in this chapter has been published recently (Lopez-Velasco et al. 2010). It presents the analysis of lettuce microflora community profiles after contamination with *E. coli* O157:H7 in refrigerated environments. Similarly to our study, authors used DGGE (Lopez-Velasco et al. 2010) to analyse changes in the lettuce microflora caused by *E. coli* and cold exposure. Authors observed that the structure of microbial communities was changing after 15 days of storage at 10°C but interestingly, the addition of *E. coli* O157:H7 caused a specific increase in epiphytic bacterial richness at 10 days, where it became the dominant microorganism (Lopez-Velasco et al. 2010). The combination of these observations with our results constitutes a strong argument to say that *E. coli*, regardless of its pathogenic status, can influence the structure of epiphytic bacterial communities.

It is somewhat surprising that GMB30, a strain that we do not believe as very adapted to persist on plants, showed such a drastic effect on resident bacterial communities, which prompts the hypothesis that this perturbation is not strain-specific and is something possibly caused by any infecting *E. coli* strain. It is also possible that GMB30 has phenotypic abilities unknown to us and conferring a particularly great ability to influence the structure of plant resident microbes. As we observed an increase in population, we earlier formulated the hypothesis that *E. coli* strain GMB30 was secreting a growth-promoting substance on plants. Good candidates for such a

substance are siderophores and it has been suggested that *Salmonella* and *E. coli* siderophores might be used by epiphytic bacteria (Brandl 2006), as this phenomenon seems to be happening frequently among leaf-associated bacteria (Loper and Buyer, 1991; Loper and Henkels, 1999). The secretion of siderophores is typically what is called in ecological terms a "public good" (Gardner and Kummerli, 2008). A strain colonising environments poor in iron will secrete iron-scavenging siderophore molecules that can be used by other individual members of the community and not necessarily the producer itself. As iron bioavailability has been described as heterogeneously distributed on leaves, it is possible that some areas in the phyllosphere are experiencing iron stress. The emigration of siderophore producers in such areas could therefore boost the growth of naturally occurring bacteria, at the expense of the contaminating bacteria.

In order to investigate this, it would be interesting to monitor community profiles on plants contaminated experimentally with other *E. coli* strains, or even non-*E. coli* bacteria such as *Salmonella*. We started to do this, but the DGGE gels were inconclusive (we could not observe any bands except for the chloroplastic DNA, even in infected samples) suggesting that we had not extracted enough DNA, and thus used enough plants per time point. However, we performed colony counts on various strains experimentally contaminating spinach grown in laboratory conditions: the high siderophore producer GMB37 (rank 1/170; PAi rank of 47/173), the low siderophore producer ECOR-63 (rank 164/170; PAi rank of 124/173) as well as *E. coli* O157:H7 strain Sakai and *Salmonella* Typhimurium., we did not observe any variation in spinach colonisation trends for any of these strains as their counts all decreased between 6 and 8 dpi (data not shown). More experiments are required to understand

the strain-specific component, if any, of the complex interactions between *E. coli* and phyllosphere resident microflora.

5.4.2. Industrial relevance and opportunities for biocontrol

There has been a growing interest over the last decades on the possibility to add exogenous biological agents, manipulate or alter non-pathogenic microbial communities in soils and on leaves in order to control or suppress the occurrence of unwanted microbes. Competitive exclusion using bacteria and fungi has proven successful to regulate insect-associated damage to crops (the most famous case being the entomopathogenic *Bacillus thuringiensis*), but it has also been examined to diminish plant disease, and more recently the spread of human pathogens on plants (Cooley et al., 2003; Hudson et al., 2009). Naturally, the first biocontrol strains to have been tested against human pathogens were the strains already proven to be efficient in controlling plant pathogens. Biocontrol *Pseudomonas fluorescens* strains (Liao and Fett, 2001; Matos et al., 2005; Fett, 2006; Liao, 2008) and different *Lactobacillus* species (Vescovo et al., 1995; Vescovo et al., 1996; Cai et al., 1997; Torriani et al., 1997) have been successfully used as antagonists against various human pathogens such as *Salmonella* (Liao and Fett, 2001; Matos et al., 2005; Fett, 2006), *Listeria monocytogenes* (Cai1997), *Staphylococcus aureus* (Vescovo, Torriani et al. 1996).

The treatment with biocontrol strains has also been investigated to limit the growth or occurrence of *E. coli* on salad leaves (Vescovo, Orsi et al. 1995) and also green peppers (Liao and Fett 2001) by competitive exclusion. In this last study, authors selected their antagonistic biocontrol strain from the analysis of the culturable natural

resident microflora of produce (Liao and Fett 2001). They screened 120 strains for the ability to inhibit the growth *in vitro* of various pathogens including *E. coli* and *Salmonella*, found 6 good candidates from *Bacillus* and *Pseudomonas* genus and confirmed their antagonistic effect on green pepper disks (Liao and Fett 2001).

In our study, although we did not thoroughly identify bacterial species corresponding to the bands appearing on denaturing gradient gels after treatment of the plants by *E. coli*, we observed that strains from the *Stenotrophomonas* and *Rahnella* genus increased in numbers when *E. coli* had infected their habitat. It somehow would mean that the presence of *E. coli* on plants triggers events leading to the growth of these bacteria on leaves. This seemingly "synergic compatibility" (i.e., the fact that the fitness of an exogenously added bacteria also benefits the fitness of the resident flora) can also be the reflection of competition for the same resources, and it remains unknown whether, if applied in high number on leaves, *Stenotrophomonas* or *Rahnella* strains would inhibit the growth of *E. coli* or not. In that respect, it is plausible to imagine naturally plant-associated strains applied as biocontrol agents, responding to *E. coli* contamination by outgrowing it and using its resources. Interestingly, it is worth mentioning that strains from the *Stenotrophomonas* and *Rahnella* genus have been successfully used in the past as biocontrol agents against fungal plant pathogens or *Xanthomonas campestris* (Kobayashi et al., 2002; El-Hendawy et al., 2005), *Ralstonia solanacearum* on potatoes (Messiha et al., 2007) or *Penicillium sp.* and *Botrytis cinerea* on apples (Calvo et al., 2007). In the light of the observations presented in this chapter, it seems plausible to further investigate the role of these bacterial species in the possible control or inhibition of *E. coli* contamination on salad leaves.

6. Conclusive remarks and perspectives

In this work, we addressed multiple questions with both academic and applied industrial views, the implications of which are detailed in this section.

- **Where does *E. coli* contaminating agricultural fields generally come from?**

We have shown that the population structure of *E. coli* contaminating plants was strongly biased in favour of phylogroup B1, and against phylogroup B2. This observation, if confirmed using plants from different other species, geographical locations, and time of isolation, could constitute an *E. coli* "ecological footprint" of plant environments, with possible implications for microbial source tracking. Indeed, a population-wide approach rather than single-strain associations seems to be the way forward, as it seems from the yet scarce literature on this topic that the association of certain environments and hosts with specific population structures is much more robust than with single-strain genotypes or markers. In **Figure 6.1**, we present the average population structures for primary and secondary environments in *E. coli*, with data taken from **Table 1.1** updated with our present analysis (additionally, see Chapter 3 section 3.6).

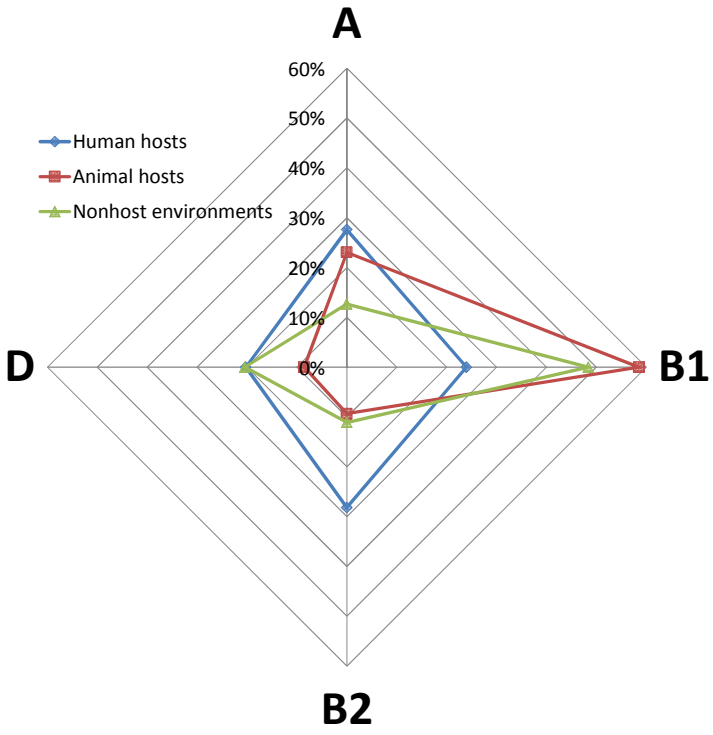


Figure 6.1. Radar plot of combined *E. coli* population structures from various environments. Data was extracted and averaged from **Table 1.1**.

Based on the overall genomic similarity, and not isolation, of plant-associated and faecal isolates, we can exclude the possibility that *E. coli* retrieved from plants come from "naturalised" populations endogenous to soil, as it was suggested earlier (Ishii et al. 2009). Unfortunately, apart from this point, we cannot answer fully the difficult question of where *E. coli* from plants comes from, but we provide a good method to do so. It would be interesting to conduct field-scale experiments over a growing season, in which the population structure of *E. coli* contaminating irrigation, wildlife or soils would be compared to the population structure on plants. Similar approaches have been preliminary investigated regarding cattle contamination of water (Carlos et al. 2010) with promising results, and it would be of prime interest to transpose this to agricultural food safety.

- **Are there specific functions or traits in plant- or nonhost-associated *E. coli*?**

A very diverse *E. coli* population, maybe broadly attributable to nonhost environments, was observed on plants and prompted the question of how host and nonhost environments can shape the population structure of *E. coli*. The simplest assumption is to consider that strains from different phylogroups possess on average different fitness-enhancing traits that modulate their abundance in various environments. In order to identify these traits, we used a combined comparative approach incorporating phylogeny and a large range of phenotypes that were reported to be involved in the colonisation and persistence of *E. coli* in various settings (see Introduction section 1.3 and Chapter 4). From our analysis, we could determine that plant and faecal isolates differed significantly on multiple levels:

(a) **plant-associated adaptation to sucrose and aromatic compounds metabolism**: We observed that GMB isolates were strikingly much better at utilising sucrose and aromatic compounds, two traits with a very strong ecological relevance in nonhost environments. Sucrose is the major carbohydrate found on plants (leaves and roots) and aromatic compounds are present at naturally very high concentrations in soils, plants and water (Diaz et al. 2001). *In vitro* competition assays are required to confirm the fitness-enhancing properties of these traits, but it is nevertheless striking to observe such an enrichment of these laterally acquired functions in plant-associated bacteria.

(b) **nonhost-associated adaptation to stress**: we extrapolated that a significantly slower metabolism on common nutrients by plant-associated isolates could be a reflection of the trade-off between stress resistance and nutritional abilities involved in the SPANC balance theory (Ferenci 2005). We unfortunately did not perform stress resistance assays but if this link is proven true, it could mean that one of the major selection pressures for high fitness in nonhost environments is the ability to better resist better various stressful conditions, reflected indirectly by a slower metabolism in GMB isolates.

(c) **phylogenetic-associated adaptation to nonhost environments**:

Based on the assumption that traits enhancing fitness in a given environment are over-represented in isolates from this environment, we calculated a "plant association index" or PAi based on the traits found to significantly discriminate GMB from ECOR. We could roughly estimate the likelihood of plant association for individual strains, and this potential was strikingly very strongly phylogenetically distributed in *E. coli*, with a high PAi values in phylogroups B1 and to some extent D, and very low PAi values in phylogroups B2 (and F), and A. We could separately confirm this distribution in plant and host isolates, indicating that this association is probably ancient in *E. coli* and was not caused by our sampling of plant-associated isolates. As presented in the conclusion of Chapter 4 (section 4.5), there is a convergence of evidence implying that *E. coli* phylogroups differ in their host association, and thus possibly in their transmission ecology, as some phylogroups are host specialists (the archetype being B2) and others are host generalists (B1) (White et al. 2011, Sims and Kim 2011). Our study contributes greatly to further this dichotomy, by showing that

host generalisation in phylogroup B1 is associated with nonhost adaptation (**Figure 6.2**).

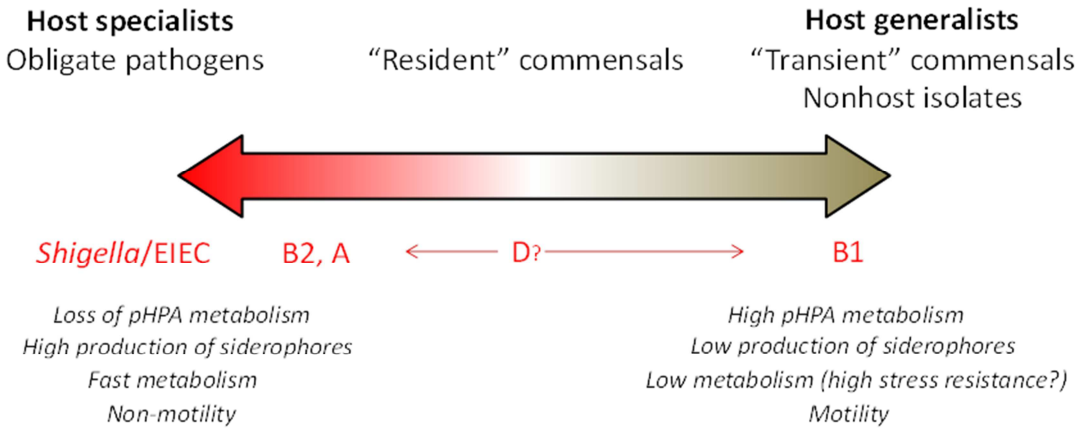


Figure 6.2. Schematic representation of the dichotomy between host specialisation and generalism in *E. coli*.

Resulting from the analysis presented in this thesis, we can present a model of host association and transmission ecology (**Figure 6.2**) in which different strains of *E. coli* range widely in their association (i.e. fitness optimum) with host and nonhost environments. In this model, the extremes of host specialisation are the obligate pathogens *Shigella* and EIEC, and host generalism and nonhost adaptation are represented by phylogroup B1 strains. It is interesting to notice that a large number of traits could potentially be associated with this ecological strategies dichotomy. It has been reported that *Shigella* had lost the ability to degrade aromatic compounds (Sabarly et al. 2011, Touchon et al. 2009), indicating that this trait is probably not important for *E. coli* strains that are strongly associated with their hosts. Accordingly, we observed a very low or nonexistent metabolism of aromatic compounds in B2 and A strains, and a high metabolism in B1, indicating that this trait could potentially be a

good marker for assessing the host adaptation status of any individual strain. There are additional minor functions that we hypothesised to be associated with specific ecological strategies (**Figure 6.2**). The high production of siderophores could be linked to adaptation to live in densely populated niches, like host intestines (see discussion in section 4.3.3), a fast metabolism and the absence of motility (see discussion in section 1.2) could additionally be associated with the intestinal environment.

- **Does *E. coli* interact with the plant resident microflora?**

We addressed this question very preliminarily in Chapter 5 of this thesis. Using DGGE, a PCR-based community fingerprinting method, we could observe that around one week after artificial contamination with *E. coli*, a surprising increase in levels of indigenous epiphytic resident bacteria occurred, possibly including bacterial species like *Stenotrophomonas sp*. and *Rahnella sp*. More research is required to investigate the mechanisms causing *E. coli* to actively (either directly or indirectly) modulate the natural epiphytic community structure.

References

Achtman, M. (2008). "Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens." Annu Rev Microbiol **62**: 53-70.

Ahmed, N., U. Dobrindt, J. Hacker and S. E. Hasnain (2008). "Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention." Nature Reviews Microbiology **6**(5): 387-394.

Alaeddinoglu, N. G. and H. P. Charles (1979). "Transfer of a gene for sucrose utilization into *Escherichia coli* K12, and consequent failure of expression of genes for D-serine utilization." J Gen Microbiol **110**(1): 47-59.

Alpert, C., J. Scheel, W. Engst, G. Loh and M. Blaut (2009). "Adaptation of protein expression by *Escherichia coli* in the gastrointestinal tract of gnotobiotic mice." Environ Microbiol **11**(4): 751-761.

Altenhoefer, A., S. Oswald, U. Sonnenborn, C. Enders, J. Schulze, J. Hacker and T. A. Oelschlaeger (2004). "The probiotic *Escherichia coli* strain Nissle 1917 interferes with invasion of human intestinal epithelial cells by different enteroinvasive bacterial pathogens." FEMS Immunol Med Microbiol **40**(3): 223-229.

Anjum, M. F., S. Lucchini, A. Thompson, J. C. Hinton and M. J. Woodward (2003). "Comparative genomic indexing reveals the phylogenomics of *Escherichia coli* pathogens." Infection and Immunity **71**(8): 4674-4683.

Arnqvist, A., A. Olsen, J. Pfeifer, D. G. Russell and S. Normark (1992). "The Crl protein activates cryptic genes for curli formation and fibronectin binding in *Escherichia coli* HB101." Mol Microbiol **6**(17): 2443-2452.

Arr, M., T. Perenyi and E. K. Novak (1970). "Sucrose and raffinose breakdown by *Escherichia coli*." Acta Microbiol Acad Sci Hung **17**(2): 117-126.

Autieri, S. M., J. J. Lins, M. P. Leatham, D. C. Laux, T. Conway and P. S. Cohen (2007). "L-fucose stimulates utilization of D-ribose by *Escherichia coli* MG1655 Δ*fucAO* and *E. coli* Nissle 1917 Δ*fucAO* mutants in the mouse intestine and in M9 minimal medium." Infection and Immunity **75**(11): 5465-5475.

Baldauf, S. L. (2003). "Phylogeny for the faint of heart: a tutorial." Trends Genet **19**(6): 345-351.

Barak, J. D., L. Gorski, P. Naraghi-Arani and A. O. Charkowski (2005). "*Salmonella enterica* virulence genes are required for bacterial attachment to plant tissue." Appl Environ Microbiol **71**(10): 5685-5691.

Barak, J. D., C. E. Jahn, D. L. Gibson and A. O. Charkowski (2007). "The role of cellulose and O-antigen capsule in the colonization of plants by *Salmonella enterica*." Mol Plant Microbe Interact **20**(9): 1083-1091.

Barak, J. D., L. C. Whitehand and A. O. Charkowski (2002). "Differences in attachment of *Salmonella enterica* serovars and *Escherichia coli* O157:H7 to alfalfa sprouts." Appl Environ Microbiol **68**(10): 4758-4763.

Barnhart, M. M., J. Lynem and M. R. Chapman (2006). "GlcNAc-6P levels modulate the expression of Curli fibers by *Escherichia coli*." Journal of Bacteriology **188**(14): 5212-5219.

Barnich, N., F. A. Carvalho, A. L. Glasser, C. Darcha, P. Jantscheff, M. Allez, H. Peeters, G. Bommelaer, P. Desreumaux, J. F. Colombel and A. Darfeuille-Michaud (2007). "CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting ileal mucosa colonization in Crohn disease." Journal of Clinical Investigation **117**(6): 1566-1574.

Beiko, R. G., T. J. Harlow and M. A. Ragan (2005). "Highways of gene sharing in prokaryotes." Proc Natl Acad Sci U S A **102**(40): 14332-14337.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Berger, C. N., D. J. Brown, R. K. Shaw, F. Minuzzi, B. Feys and G. Frankel (2011). "*Salmonella enterica* strains belonging to O serogroup 1,3,19 induce chlorosis and wilting of *Arabidopsis thaliana* leaves." Environ Microbiol **13**(5): 1299-1308.

Berger, C. N., R. K. Shaw, D. J. Brown, H. Mather, S. Clare, G. Dougan, M. J. Pallen and G. Frankel (2009). "Interaction of *Salmonella enterica* with basil and other salad leaves." ISME J **3**(2): 261-265.

Bergholz, P. W., J. D. Noar and D. H. Buckley (2011). "Environmental patterns are imposed on the population structure of *Escherichia coli* after fecal deposition." Appl Environ Microbiol **77**(1): 211-219.

Bergthorsson, U. and H. Ochman (1998). "Distribution of chromosome length variation in natural isolates of *Escherichia coli*." Molecular Biology and Evolution **15**(1): 6-16.

Bermudez, M. and T. C. Hazen (1988). "Phenotypic and genotypic comparison of *Escherichia coli* from pristine tropical waters." Appl Environ Microbiol **54**(4): 979-983.

Bertin, Y., J. P. Girardeau, F. Chaucheyras-Durand, B. Lyan, E. Pujos-Guillot, J. Harel and C. Martin (2011). "Enterohaemorrhagic *Escherichia coli* gains a competitive advantage by using ethanolamine as a nitrogen source in the bovine intestinal content." Environ Microbiol **13**(2): 365-377.

Beuchat, L. R. (2002). "Ecological factors influencing survival and growth of human pathogens on raw fruits and vegetables." Microbes and Infection **4**(4): 413-423.

Beuchat, L. R., T. E. Ward and C. A. Pettigrew (2001). "Comparison of chlorine and a prototype produce wash product for effectiveness in killing *Salmonella* and *Escherichia coli* O157:H7 on alfalfa seeds." J Food Prot **64**(2): 152-158.

Bhagwat, A. A., L. Chan, R. Han, J. Tan, M. Kothary, J. Jean-Gilles and B. D. Tall (2005). "Characterization of enterohemorrhagic *Escherichia coli* strains based on acid resistance phenotypes." Infect Immun **73**(8): 4993-5003.

Bianco, C., E. Imperlini, R. Calogero, B. Senatore, A. Amoresano, A. Carpentieri, P. Pucci and R. Defez (2006). "Indole-3-acetic acid improves *Escherichia coli*'s defences to stress." Arch Microbiol **185**(5): 373-382.

Bibbal, D., V. Dupouy, M. F. Prere, P. L. Toutain and A. Bousquet-Melou (2009). "Relatedness of *Escherichia coli* strains with different susceptibility phenotypes isolated from swine feces during ampicillin treatment." Appl Environ Microbiol **75**(10): 2999-3006.

Bingen-Bidois, M., O. Clermont, S. Bonacorsi, M. Terki, N. Brahimi, C. Loukil, D. Barraud and E. Bingen (2002). "Phylogenetic analysis and prevalence of urosepsis strains of *Escherichia coli* bearing pathogenicity island-like domains." Infection and Immunity **70**(6): 3216-3226.

Bochner, B. R. (2009). "Global phenotypic characterization of bacteria." FEMS Microbiol Rev **33**(1): 191-205.

Bochner, B. R. and M. A. Savageau (1977). "Generalized indicator plate for genetic, metabolic, and taxonomic studies with microorganisms." Appl Environ Microbiol **33**(2): 434-444.

Booijink, C. C., J. Boekhorst, E. G. Zoetendal, H. Smidt, M. Kleerebezem and W. M. de Vos (2010). "Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed." Appl Environ Microbiol **76**(16): 5533-5540.

Bossio, D. A. and K. M. Scow (1995). "Impact of carbon and flooding on the metabolic diversity of microbial communities in soils." Appl Environ Microbiol **61**(11): 4043-4050.

Brandl, M. T. (2006). "Fitness of human enteric pathogens on plants and implications for food safety." Annu Rev Phytopathol **44**: 367-392.

Brandl, M. T. and R. Amundson (2008). "Leaf age as a risk factor in contamination of lettuce with *Escherichia coli* O157:H7 and *Salmonella enterica*." Appl Environ Microbiol **74**(8): 2298-2306.

Brennan, F. P., F. Abram, F. A. Chinalia, K. G. Richards and V. O'Flaherty (2010). "Characterization of environmentally persistent *Escherichia coli* isolates leached from an Irish soil." Appl Environ Microbiol **76**(7): 2175-2180.

Brennan, F. P., V. O'Flaherty, G. Kramers, J. Grant and K. G. Richards (2010). "Long-term persistence and leaching of *Escherichia coli* in temperate maritime soils." Appl Environ Microbiol **76**(5): 1449-1455.

Buitenhuis, B., C. M. Rontved, S. M. Edwards, K. L. Ingvartsen and P. Sorensen (2011). "In depth analysis of genes and pathways of the mammary gland involved in the pathogenesis of bovine *Escherichia coli*-mastitis." BMC Genomics **12**: 130.

Bunnik, E. M., L. C. Swenson, D. Edo-Matas, W. Huang, W. Dong, A. Frantzell, C. J. Petropoulos, E. Coakley, H. Schuitemaker, P. R. Harrigan and A. B. van 't Wout (2011). "Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra-deep pyrosequencing." Plos Pathogens **7**(6): e1002106.

Byappanahalli, M. N., R. L. Whitman, D. A. Shively, M. J. Sadowsky and S. Ishii (2006). "Population structure, persistence, and seasonality of autochthonous *Escherichia coli* in temperate, coastal forest soil from a Great Lakes watershed." Environ Microbiol **8**(3): 504-513.

Cai, Y., L. K. Ng and J. M. Farber (1997). "Isolation and characterization of nisin-producing *Lactococcus lactis subsp. lactis* from bean-sprouts." J Appl Microbiol **83**(4): 499-507.

Caldwell, K. N., G. L. Anderson, P. L. Williams and L. R. Beuchat (2003). "Attraction of a free-living nematode, *Caenorhabditis elegans*, to foodborne pathogenic bacteria and its potential as a vector of *Salmonella poona* for preharvest contamination of cantaloupe." J Food Prot **66**(11): 1964-1971.

Calvin, L., H. H. Jensen and J. Liang (2009). The Economics of Food Safety: The 2006 Foodborne Illness Outbreak Linked to Spinach. Microbial Safety of Fresh Produce, Wiley-Blackwell**:** 399-417.

Calvo, J., V. Calvente, M. E. de Orellano, D. Benuzzi and M. I. Sanz de Tosetti (2007). "Biological control of postharvest spoilage caused by *Penicillium expansum* and *Botrytis cinerea* in apple by using the bacterium *Rahnella aquatilis*." Int J Food Microbiol **113**(3): 251-257.

Caponigro, V., M. Ventura, I. Chiancone, L. Amato, E. Parente and F. Piro (2010). "Variation of microbial load and visual quality of ready-to-eat salads by vegetable type, season, processor and retailer." Food Microbiol **27**(8): 1071-1077.

Carlos, C., M. M. Pires, N. C. Stoppe, E. M. Hachich, M. I. Sato, T. A. Gomes, L. A. Amaral and L. M. Ottoboni (2010). "*Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination." BMC Microbiology **10**: 161.

Caugant, D. A., B. R. Levin and R. K. Selander (1981). "Genetic diversity and temporal variation in the *E. coli* population of a human host." Genetics **98**(3): 467-490.

Caza, M., F. Lepine, S. Milot and C. M. Dozois (2008). "Specific roles of the *iroBCDEN* genes in virulence of an avian pathogenic *Escherichia coli* O78 strain and in production of salmochelins." Infection and Immunity **76**(8): 3539-3549.

Chang, D. E., D. J. Smalley, D. L. Tucker, M. P. Leatham, W. E. Norris, S. J. Stevenson, A. B. Anderson, J. E. Grissom, D. C. Laux, P. S. Cohen and T. Conway (2004). "Carbon nutrition of *Escherichia coli* in the mouse intestine." Proc Natl Acad Sci U S A **101**(19): 7427-7432.

Chao, A. (1984). "Nonparametric Estimation of the Number of Classes in a Population." Scandinavian Journal of Statistics **11**(4): 265-270.

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

Clermont, O., D. M. Gordon, S. Brisse, S. T. Walk and E. Denamur (2011). "Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence." Environ Microbiol.

Clermont, O., M. Lescat, C. L. O'Brien, D. M. Gordon, O. Tenaillon and E. Denamur (2008). "Evidence for a human-specific *Escherichia coli* clone." Environ Microbiol **10**(4): 1000-1006.

Clermont, O., M. Olier, C. Hoede, L. Diancourt, S. Brisse, M. Keroudean, J. Glodt, B. Picard, E. Oswald and E. Denamur (2011). "Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds." Infection Genetics and Evolution **11**(3): 654-662.

Colwell, R. K., C. X. Mao and J. Chang (2004). "Interpolating, Extrapolating, and Comparing Incidence-Based Species Accumulation Curves." Ecology **85**(10): 2717-2727.

Cooke, E. M., I. G. Hettiaratchy and A. C. Buck (1972). "Fate of ingested *Escherichia coli* in normal persons." J Med Microbiol **5**(3): 361-369.

Cooley, M. B., D. Chao and R. E. Mandrell (2006). "*Escherichia coli* O157:H7 survival and growth on lettuce is altered by the presence of epiphytic bacteria." J Food Prot **69**(10): 2329-2335.

Cooley, M. B., W. G. Miller and R. E. Mandrell (2003). "Colonization of *Arabidopsis thaliana* with *Salmonella enterica* and enterohemorrhagic *Escherichia coli* O157:H7 and competition by *Enterobacter asburiae*." Appl Environ Microbiol **69**(8): 4915-4926.

Cooper, V. S. and R. E. Lenski (2000). "The population genetics of ecological specialization in evolving *Escherichia coli* populations." Nature **407**(6805): 736-739.

Cornelis, P. (2010). "Iron uptake and metabolism in pseudomonads." Appl Microbiol Biotechnol **86**(6): 1637-1645.

Croxen, M. A. and B. B. Finlay (2010). "Molecular mechanisms of *Escherichia coli* pathogenicity." Nature Reviews Microbiology **8**(1): 26-38.

Currie, B. J., D. Gal, M. Mayo, L. Ward, D. Godoy, B. G. Spratt and J. J. LiPuma (2007). "Using BOX-PCR to exclude a clonal outbreak of melioidosis." BMC Infect Dis **7**: 68.

Czeczulin, J. R., S. Balepur, S. Hicks, A. Phillips, R. Hall, M. H. Kothary, F. Navarro-Garcia and J. P. Nataro (1997). "Aggregative adherence fimbria II, a second fimbrial antigen mediating aggregative adherence in enteroaggregative *Escherichia coli*." Infection and Immunity **65**(10): 4135-4145.

Darling, A. E., I. Miklos and M. A. Ragan (2008). "Dynamics of genome rearrangement in bacterial populations." Plos Genetics **4**(7): e1000128.

de Araujo, J. C. and R. P. Schneider (2008). "DGGE with genomic DNA: suitable for detection of numerically important organisms but not for identification of the most abundant organisms." Water Res **42**(20): 5002-5010.

de Muinck, E. J., T. Øien, O. Storrø, R. Johnsen, N. C. Stenseth, K. S. Rønningen and K. Rudi (2011). "Diversity, transmission and persistence of *Escherichia coli* in a cohort of mothers and their infants." Environ Microbiol Rep **3**(3): 352-359.

De Paepe, M., V. Gaboriau-Routhiau, D. Rainteau, S. Rakotobe, F. Taddei and N. Cerf-Bensussan (2011). "Trade-off between bile resistance and nutritional competence drives *Escherichia coli* diversification in the mouse gut." Plos Genetics **7**(6): e1002107.

de Vienne, D. M., T. Giraud and O. C. Martin (2007). "A congruence index for testing topological similarity between trees." Bioinformatics **23**(23): 3119-3124.

Deering, A. J., L. J. Mauer and R. E. Pruitt (2011). "Internalization of *E. coli* O157:H7 and *Salmonella spp.* in plants: A review." Food Research International(0).

Deszo, E. L., S. M. Steenbergen, D. I. Freedberg and E. R. Vimr (2005). "*Escherichia coli* K1 polysialic acid O-acetyltransferase gene, *neuO*, and the mechanism of capsule form variation involving a mobile contingency locus." Proc Natl Acad Sci U S A **102**(15): 5564-5569.

Di Giovanni, G. D., L. S. Watrud, R. J. Seidler and F. Widmer (1999). "Comparison of Parental and Transgenic Alfalfa Rhizosphere Bacterial Communities Using Biolog GN Metabolic Fingerprinting and Enterobacterial Repetitive Intergenic Consensus Sequence-PCR (ERIC-PCR)." Microb Ecol **37**(2): 129-139.

Diallo, S., A. Crepin, C. Barbey, N. Orange, J. F. Burini and X. Latour (2011). "Mechanisms and recent advances in biological control mediated through the potato rhizosphere." FEMS Microbiol Ecol **75**(3): 351-364.

Diaz, E., A. Ferrandez, M. A. Prieto and J. L. Garcia (2001). "Biodegradation of aromatic compounds by *Escherichia coli*." Microbiol Mol Biol Rev **65**(4): 523-569, table of contents.

Didelot, X., M. Barker, D. Falush and F. G. Priest (2009). "Evolution of pathogenicity in the *Bacillus cereus* group." Syst Appl Microbiol **32**(2): 81-90.

Didelot, X., R. Bowden, T. Street, T. Golubchik, C. Spencer, G. McVean, V. Sangal, M. F. Anjum, M. Achtman, D. Falush and P. Donnelly (2011). "Recombination and population structure in *Salmonella enterica*." Plos Genetics **7**(7): e1002191.

Didelot, X. and D. Falush (2007). "Inference of bacterial microevolution using multilocus sequence data." Genetics **175**(3): 1251-1266.

Didelot, X. and M. C. Maiden (2010). "Impact of recombination on bacterial evolution." Trends Microbiol **18**(7): 315-322.

Dobbin, H. S., C. J. Hovde, C. J. Williams and S. A. Minnich (2006). "The *Escherichia coli* O157 flagellar regulatory gene *flhC* and not the flagellin gene *fliC* impacts colonization of cattle." Infection and Immunity **74**(5): 2894-2905.

Dobrindt, U. and J. Hacker (2001). "Whole genome plasticity in pathogenic bacteria." Curr Opin Microbiol **4**(5): 550-557.

Dong, Y., A. L. Iniguez, B. M. Ahmer and E. W. Triplett (2003). "Kinetics and strain specificity of rhizosphere and endophytic colonization by enteric bacteria on seedlings of *Medicago sativa* and *Medicago truncatula*." Appl Environ Microbiol **69**(3): 1783-1790.

Donohue-Rolfe, A., D. W. Acheson and G. T. Keusch (1991). "Shiga toxin: purification, structure, and function." Rev Infect Dis **13 Suppl 4**: S293-297.

Dufour, D., P. Germon, E. Brusseaux, Y. Le Roux and A. Dary (2011). "First evidence of the presence of genomic islands in *Escherichia coli* P4, a mammary pathogen frequently used to induce experimental mastitis." J Dairy Sci **94**(6): 2779-2793.

Duriez, P., O. Clermont, S. Bonacorsi, E. Bingen, A. Chaventre, J. Elion, B. Picard and E. Denamur (2001). "Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations." Microbiology **147**(Pt 6): 1671-1676.

Dykhuizen, D. E. and L. Green (1991). "Recombination in *Escherichia coli* and the definition of biological species." Journal of Bacteriology **173**(22): 7257-7268.

El-Hendawy, H. H., M. E. Osman and N. M. Sorour (2005). "Biological control of bacterial spot of tomato caused by *Xanthomonas campestris pv. vesicatoria* by *Rahnella aquatilis*." Microbiol Res **160**(4): 343-352.

Escobar-Paramo, P., O. Clermont, A. B. Blanc-Potard, H. Bui, C. Le Bouguenec and E. Denamur (2004). "A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*." Molecular Biology and Evolution **21**(6): 1085-1094.

Espinosa-Urgel, M. (2004). "Plant-associated *Pseudomonas* populations: molecular biology, DNA dynamics, and gene transfer." Plasmid **52**(3): 139-150.

Evans, D. G., D. J. Evans, Jr. and W. Tjoa (1977). "Hemagglutination of human group A erythrocytes by enterotoxigenic *Escherichia coli* isolated from adults with diarrhea: correlation with colonization factor." Infection and Immunity **18**(2): 330-337.

Fabich, A. J., S. A. Jones, F. Z. Chowdhury, A. Cernosek, A. Anderson, D. Smalley, J. W. McHargue, G. A. Hightower, J. T. Smith, S. M. Autieri, M. P. Leatham, J. J. Lins, R. L. Allen, D. C. Laux, P. S. Cohen and T. Conway (2008). "Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine." Infection and Immunity **76**(3): 1143-1152.

Ferenci, T. (2005). "Maintaining a healthy SPANC balance through regulatory and mutational adaptation." Mol Microbiol **57**(1): 1-8.

Ferenci, T., H. F. Galbiati, T. Betteridge, K. Phan and B. Spira (2011). "The constancy of global regulation across a species: the concentrations of ppGpp and RpoS are strain-specific in *Escherichia coli*." BMC Microbiol **11**: 62.
Fett, W. F. (2006). "Inhibition of *Salmonella enterica* by plant-associated pseudomonads in vitro and on sprouting alfalfa seed." J Food Prot **69**(4): 719-728.
Fierer, N. and R. B. Jackson (2006). "The diversity and biogeography of soil bacterial communities." Proc Natl Acad Sci U S A **103**(3): 626-631.
Finkel, S. E. and R. Kolter (2001). "DNA as a nutrient: novel role for bacterial competence gene homologs." Journal of Bacteriology **183**(21): 6288-6293.
Fischer, S. G. and L. S. Lerman (1979). "Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis." Cell **16**(1): 191-200.
Fischer, S. G. and L. S. Lerman (1980). "Separation of random fragments of DNA according to properties of their sequences." Proc Natl Acad Sci U S A **77**(8): 4420-4424.
Frank, S. A. and P. Schmid-Hempel (2008). "Mechanisms of pathogenesis and the evolution of parasite virulence." J Evol Biol **21**(2): 396-404.
Friedmann, H. C. (2006). "Escherich and *Escherichia*." Adv Appl Microbiol **60**: 133-196.
Fukushima, H., T. Hashizume, Y. Morita, J. Tanaka, K. Azuma, Y. Mizumoto, M. Kaneno, M. Matsuura, K. Konma and T. Kitani (1999). "Clinical experiences in Sakai City Hospital during the massive outbreak of enterohemorrhagic *Escherichia coli* O157 infections in Sakai City, 1996." Pediatr Int **41**(2): 213-217.
Garbeva, P., J. A. van Veen and J. D. van Elsas (2004). "Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness." Annu Rev Phytopathol **42**: 243-270.
Gardner, A. and R. Kummerli (2008). "Social evolution: this microbe will self-destruct." Curr Biol **18**(21): R1021-1023.
Garenaux, A., M. Caza and C. M. Dozois (2011). "The Ins and Outs of siderophore mediated iron uptake by extra-intestinal pathogenic *Escherichia coli*." Vet Microbiol **153**(1-2): 89-98.
Garmendia, J., G. Frankel and V. F. Crepin (2005). "Enteropathogenic and enterohemorrhagic *Escherichia coli* infections: translocation, translocation, translocation." Infection and Immunity **73**(5): 2573-2585.
Gauger, E. J., M. P. Leatham, R. Mercado-Lubo, D. C. Laux, T. Conway and P. S. Cohen (2007). "Role of motility and the *flhDC* Operon in *Escherichia coli* MG1655 colonization of the mouse intestine." Infection and Immunity **75**(7): 3315-3324.
Gerstel, U. and U. Romling (2001). "Oxygen tension and nutrient starvation are major signals that regulate *agfD* promoter activity and expression of the multicellular morphotype in *Salmonella typhimurium*." Environ Microbiol **3**(10): 638-648.
Gevers, D., K. Vandepoele, C. Simillon and Y. Van de Peer (2004). "Gene duplication and biased functional retention of paralogs in bacterial genomes." Trends Microbiol **12**(4): 148-154.
Gibson, D. L., A. P. White, S. D. Snyder, S. Martin, C. Heiss, P. Azadi, M. Surette and W. W. Kay (2006). "*Salmonella* produces an O-antigen capsule regulated by AgfD and important for environmental persistence." Journal of Bacteriology **188**(22): 7722-7730.
Gordon, D. M., S. Bauer and J. R. Johnson (2002). "The genetic structure of *Escherichia coli* populations in primary and secondary habitats." Microbiology **148**(Pt 5): 1513-1522.
Gordon, D. M., O. Clermont, H. Tolley and E. Denamur (2008). "Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method." Environ Microbiol **10**(10): 2484-2496.
Gordon, D. M. and A. Cowling (2003). "The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects." Microbiology **149**(Pt 12): 3575-3586.
Goto, D. K. and T. Yan (2011). "Genotypic diversity of *Escherichia coli* in the water and soil of tropical watersheds in Hawaii." Appl Environ Microbiol **77**(12): 3988-3997.

Gourabathini, P., M. T. Brandl, K. S. Redding, J. H. Gunderson and S. G. Berk (2008). "Interactions between food-borne pathogens and protozoa isolated from lettuce and spinach." Appl Environ Microbiol **74**(8): 2518-2525.

Grimaldi, D., S. Bonacorsi, H. Roussel, B. Zuber, H. Poupet, J. D. Chiche, C. Poyart and J. P. Mira (2010). "Unusual "flesh-eating" strain of *Escherichia coli*." J Clin Microbiol **48**(10): 3794-3796.

Grunwald, N. J. and E. M. Goss (2011). "Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches." Annu Rev Phytopathol **49**: 249-267.

Guo, X., J. Chen, R. E. Brackett and L. R. Beuchat (2001). "Survival of salmonellae on and in tomato plants from the time of inoculation at flowering and early stages of fruit development through fruit ripening." Appl Environ Microbiol **67**(10): 4760-4764.

Hacker, J. and E. Carniel (2001). "Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes." EMBO Rep **2**(5): 376-381.

Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucl Acids Symp Ser **41**: 95-98.

Hancock, V., A. S. Seshasayee, D. W. Ussery, N. M. Luscombe and P. Klemm (2008). "Transcriptomics and adaptive genomics of the asymptomatic bacteriuria *Escherichia coli* strain 83972." Mol Genet Genomics **279**(5): 523-534.

Hartl, D. L. and D. E. Dykhuizen (1984). "The population genetics of *Escherichia coli*." Annu Rev Genet **18**: 31-68.

Heaton, J. C. and K. Jones (2008). "Microbial contamination of fruit and vegetables and the behaviour of enteropathogens in the phyllosphere: a review." J Appl Microbiol **104**(3): 613-626.

Hedberg, C. W., F. J. Angulo, K. E. White, C. W. Langkop, W. L. Schell, M. G. Stobierski, A. Schuchat, J. M. Besser, S. Dietrich, L. Helsel, P. M. Griffin, J. W. McFarland and M. T. Osterholm (1999). "Outbreaks of salmonellosis associated with eating uncooked tomatoes: implications for public health. The Investigation Team." Epidemiol Infect **122**(3): 385-393.

Herias, M. V., T. Midtvedt, L. A. Hanson and A. E. Wold (1995). "Role of *Escherichia coli* P fimbriae in intestinal colonization in gnotobiotic rats." Infection and Immunity **63**(12): 4781-4789.

Herias, M. V., T. Midtvedt, L. A. Hanson and A. E. Wold (1997). "*Escherichia coli* K5 capsule expression enhances colonization of the large intestine in the gnotobiotic rat." Infection and Immunity **65**(2): 531-536.

Herzer, P. J., S. Inouye, M. Inouye and T. S. Whittam (1990). "Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*." Journal of Bacteriology **172**(11): 6175-6181.

Hong, S., J. Bunge, C. Leslin, S. Jeon and S. S. Epstein (2009). "Polymerase chain reaction primers miss half of rRNA microbial diversity." ISME J **3**(12): 1365-1373.

Hudson, J. A., C. Billington and L. McIntyre (2009). Biological Control of Human Pathogens on Produce. Microbial Safety of Fresh Produce, Wiley-Blackwell**:** 205-224.

Hughes, D. (2000). "Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes." Genome Biol **1**(6): REVIEWS0006.

Hughes, J. B. and J. J. Hellmann (2005). "The application of rarefaction techniques to molecular inventories of microbial diversity." Methods Enzymol **397**: 292-308.

Huson, D. H. (1998). "SplitsTree: analyzing and visualizing evolutionary data." Bioinformatics **14**(1): 68-73.

Ibarz Pavon, A. B. and M. C. Maiden (2009). "Multilocus sequence typing." Methods Mol Biol **551**: 129-140.

Ibekwe, A. M., C. M. Grieve and C. H. Yang (2007). "Survival of *Escherichia coli* O157:H7 in soil and on lettuce after soil fumigation." Can J Microbiol **53**(5): 623-635.

Ibenyassine, K., R. A. Mhand, Y. Karamoko, B. Anajjar, M. M. Chouibani and M. Ennaji (2007). "Bacterial pathogens recovered from vegetables irrigated by wastewater in Morocco." J Environ Health **69**(10): 47-51.

Ihssen, J., E. Grasselli, C. Bassin, P. Francois, J. C. Piffaretti, W. Koster, J. Schrenzel and T. Egli (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 **153**(Pt 7): 2052-2066.

Ilic, S., J. Odomeru and J. T. LeJeune (2008). "Coliforms and prevalence of *Escherichia coli* and foodborne pathogens on minimally processed spinach in two packing plants." J Food Prot **71**(12): 2398-2403.

Ingle, D. J., O. Clermont, D. Skurnik, E. Denamur, S. T. Walk and D. M. Gordon (2011). "Biofilm formation by and thermal niche and virulence characteristics of *Escherichia spp*." Appl Environ Microbiol **77**(8): 2695-2700.

Ishii, S., D. L. Hansen, R. E. Hicks and M. J. Sadowsky (2007). "Beach sand and sediments are temporal sinks and sources of *Escherichia coli* in Lake Superior." Environmental Science & Technology **41**(7): 2203-2209.

Ishii, S., W. B. Ksoll, R. E. Hicks and M. J. Sadowsky (2006). "Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior watersheds." Appl Environ Microbiol **72**(1): 612-621.

Ishii, S. and M. J. Sadowsky (2008). "*Escherichia coli* in the Environment: Implications for Water Quality and Human Health." Microbes Environ **23**(2): 101-108.

Ishii, S. and M. J. Sadowsky (2009). "Applications of the rep-PCR DNA fingerprinting technique to study microbial diversity, ecology and evolution." Environ Microbiol **11**(4): 733-740.

Ishii, S., T. Yan, H. Vu, D. L. Hansen, R. E. Hicks and M. J. Sadowsky (2010). "Factors controlling long-term survival and growth of naturalized *Escherichia coli* populations in temperate field soils." Microbes Environ **25**(1): 8-14.

Islam, M., M. P. Doyle, S. C. Phatak, P. Millner and X. Jiang (2004). "Persistence of enterohemorrhagic *Escherichia coli* O157:H7 in soil and on leaf lettuce and parsley grown in fields treated with contaminated manure composts or irrigation water." J Food Prot **67**(7): 1365-1370.

Itoh, Y., Y. Sugita-Konishi, F. Kasuga, M. Iwaki, Y. Hara-Kudo, N. Saito, Y. Noguchi, H. Konuma and S. Kumagai (1998). "Enterohemorrhagic *Escherichia coli* O157:H7 present in radish sprouts." Appl Environ Microbiol **64**(4): 1532-1535.

Jahreis, K., L. Bentler, J. Bockmann, S. Hans, A. Meyer, J. Siepelmeyer and J. W. Lengeler (2002). "Adaptation of sucrose metabolism in the *Escherichia coli* wild-type strain EC3132." J Bacteriol **184**(19): 5307-5316.

Janisiewicz, W. J., W. S. Conway, M. W. Brown, G. M. Sapers, P. Fratamico and R. L. Buchanan (1999). "Fate of *Escherichia coli* O157:H7 on fresh-cut apple tissue and its potential for transmission by fruit flies." Appl Environ Microbiol **65**(1): 1-5.

Jaureguy, F., L. Landraud, V. Passet, L. Diancourt, E. Frapy, G. Guigon, E. Carbonnelle, O. Lortholary, O. Clermont, E. Denamur, B. Picard, X. Nassif and S. Brisse (2008). "Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains." BMC Genomics **9**: -.

Jeter, C. and A. G. Matthysse (2005). "Characterization of the binding of diarrheagenic strains of *E. coli* to plant surfaces and the role of curli in the interaction of the bacteria with alfalfa sprouts." Mol Plant Microbe Interact **18**(11): 1235-1242.

Johnson, J. R., O. Clermont, M. Menard, M. A. Kuskowski, B. Picard and E. Denamur (2006). "Experimental mouse lethality of *Escherichia coli* isolates, in relation to accessory traits, phylogenetic group, and ecological source." Journal of Infectious Diseases **194**(8): 1141-1150.

Johnson, K. B., V. O. Stockwell, D. M. Burgett, D. Sugar and J. E. Loper (1993). "Dispersal of *Erwinia amylovora* and *Pseudomonas fluorescens* by honey bees from hives to apple and pear blossoms." Phytopathology **83**(5): 478-484.

Jones, S. A., F. Z. Chowdhury, A. J. Fabich, A. Anderson, D. M. Schreiner, A. L. House, S. M. Autieri, M. P. Leatham, J. J. Lins, M. Jorgensen, P. S. Cohen and T. Conway (2007). "Respiration of *Escherichia coli* in the mouse intestine." Infection and Immunity **75**(10): 4891-4899.

Jones, S. A., M. Jorgensen, F. Z. Chowdhury, R. Rodgers, J. Hartline, M. P. Leatham, C. Struve, K. A. Krogfelt, P. S. Cohen and T. Conway (2008). "Glycogen and maltose utilization by *Escherichia coli* O157:H7 in the mouse intestine." Infection and Immunity **76**(6): 2531-2540.

Jordan, I. K., K. S. Makarova, Y. I. Wolf and E. V. Koonin (2001). "Gene conversions in genes encoding outer-membrane proteins in *H. pylori* and *C. pneumoniae*." Trends Genet **17**(1): 7-10.

Joshi, F., G. Archana and A. Desai (2006). "Siderophore cross-utilization amongst rhizospheric bacteria and the role of their differential affinities for Fe3+ on growth stimulation under iron-limited conditions." Curr Microbiol **53**(2): 141-147.

Joshi, F. R., S. P. Kholiya, G. Archana and A. J. Desai (2008). "Siderophore cross-utilization amongst nodule isolates of the cowpea miscellany group and its effect on plant growth in the presence of antagonistic organisms." Microbiol Res **163**(5): 564-570.

Joyner, D. C. and S. E. Lindow (2000). "Heterogeneity of iron bioavailability on plants assessed with a whole-cell GFP-based bacterial biosensor." Microbiology **146 (Pt 10)**: 2435-2445.

Jurkevitch, E., Y. Hadar and Y. Chen (1992). "Differential siderophore utilization and iron uptake by soil and rhizosphere bacteria." Appl Environ Microbiol **58**(1): 119-124.

Kadivar, H. and A. E. Stapleton (2003). "Ultraviolet radiation alters maize phyllosphere bacterial diversity." Microb Ecol **45**(4): 353-361.

Kang, Y. J., J. Cheng, L. J. Mei, J. Hu, Z. Piao and S. X. Yin (2010). "Multiple copies of 16S rRNA gene affect the restriction patterns and DGGE profile as revealed by analysis of genome database." Mikrobiologiia **79**(5): 664-671.

Kaper, J. B., J. P. Nataro and H. L. Mobley (2004). "Pathogenic *Escherichia coli*." Nature Reviews Microbiology **2**(2): 123-140.

Kenney, S. J., G. L. Anderson, P. L. Williams, P. D. Millner and L. R. Beuchat (2005). "Persistence of *Escherichia coli* O157:H7, *Salmonella Newport*, and *Salmonella Poona* in the gut of a free-living nematode, *Caenorhabditis elegans*, and transmission to progeny and uninfected nematodes." Int J Food Microbiol **101**(2): 227-236.

Kenney, S. J., G. L. Anderson, P. L. Williams, P. D. Millner and L. R. Beuchat (2006). "Migration of *Caenorhabditis elegans* to manure and manure compost and potential for transport of *Salmonella newport* to fruits and vegetables." Int J Food Microbiol **106**(1): 61-68.

Kim, C. C., E. A. Joyce, K. Chan and S. Falkow (2002). "Improved analytical methods for microarray-based genome-composition analysis." Genome Biol **3**(11): RESEARCH0065.

King, T., A. Ishihama, A. Kori and T. Ferenci (2004). "A regulatory trade-off as a source of strain variation in the species *Escherichia coli*." J Bacteriol **186**(17): 5614-5620.

Kist, M. (1986). "[Who discovered *Campylobacter jejuni/coli*? A review of hitherto disregarded literature]." Zentralbl Bakteriol Mikrobiol Hyg A **261**(2): 177-186.

Kniskern, J. M., M. B. Traw and J. Bergelson (2007). "Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on *Arabidopsis thaliana*." Mol Plant Microbe Interact **20**(12): 1512-1522.

Kobayashi, D. Y., R. M. Reedy, J. Bick and P. V. Oudemans (2002). "Characterization of a chitinase gene from *Stenotrophomonas maltophilia* strain 34S1 and its involvement in biological control." Appl Environ Microbiol **68**(3): 1047-1054.

Koch, A. (1987). Why *Escherichia coli* should be renamed *Escherichia ilei*. Phosphate metabolism and cellular regulation in microorganisms. F. G. R. A. Torriani, S. Silver, A. Wright, and E. Yagil (ed.). Washington, D.C., ASM Press.

Konstantinidis, K. T., A. Ramette and J. M. Tiedje (2006). "Toward a more robust assessment of intraspecies diversity, using fewer genetic markers." Appl Environ Microbiol **72**(11): 7286-7293.

Kragelund, L., C. Hosbond and O. Nybroe (1997). "Distribution of metabolic activity and phosphate starvation response of lux-tagged *Pseudomonas fluorescens* reporter bacteria in the barley rhizosphere." Appl Environ Microbiol **63**(12): 4920-4928.

Kreader, C. A. (1996). "Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein." Appl Environ Microbiol **62**(3): 1102-1106.

Kroupitski, Y., D. Golberg, E. Belausov, R. Pinto, D. Swartzberg, D. Granot and S. Sela (2009). "Internalization of *Salmonella enterica* in leaves is induced by light and involves chemotaxis and penetration through open stomata." Appl Environ Microbiol **75**(19): 6076-6086.

Kulasekara, B. R., M. Jacobs, Y. Zhou, Z. Wu, E. Sims, C. Saenphimmachak, L. Rohmer, J. M. Ritchie, M. Radey, M. McKevitt, T. L. Freeman, H. Hayden, E. Haugen, W. Gillett, C. Fong, J. Chang, V. Beskhlebnaya, M. K. Waldor, M. Samadpour, T. S. Whittam, R. Kaul, M. Brittnacher and S. I. Miller (2009). "Analysis of the genome of the *Escherichia coli* O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence." Infection and Immunity **77**(9): 3713-3721.

Kuo, C. H. and H. Ochman (2010). "The extinction dynamics of bacterial pseudogenes." Plos Genetics **6**(8).

Kutter, S., A. Hartmann and M. Schmid (2006). "Colonization of barley (*Hordeum vulgare*) with *Salmonella enterica* and *Listeria spp.*" FEMS Microbiol Ecol **56**(2): 262-271.

Kyle, J. L., C. T. Parker, D. Goudeau and M. T. Brandl (2010). "Transcriptome analysis of *Escherichia coli* O157:H7 exposed to lysates of lettuce leaves." Appl Environ Microbiol **76**(5): 1375-1387.

Lan, R. and P. R. Reeves (2002). "*Escherichia coli* in disguise: molecular origins of *Shigella*." Microbes and Infection **4**(11): 1125-1132.

Lan, R., G. Stevenson and P. R. Reeves (2003). "Comparison of two major forms of the *Shigella* virulence plasmid pINV: positive selection is a major force driving the divergence." Infection and Immunity **71**(11): 6298-6306.

Landini, P., G. Jubelin and C. Dorel-Flamant (2006). The Molecular Genetics of Bioadhesion and Biofilm Formation. Biological Adhesives. J. A. C. e. A. N. Smith. Berlin Heidelberg, Springer-Verlag**:** 21-35.

Lapidot, A., U. Romling and S. Yaron (2006). "Biofilm formation and the survival of *Salmonella* Typhimurium on parsley." Int J Food Microbiol **109**(3): 229-233.

Law, D. (2000). "Virulence factors of *Escherichia coli* O157 and other Shiga toxin-producing *E. coli*." J Appl Microbiol **88**(5): 729-745.

Le Bouguenec, C. (2005). "Adhesins and invasins of pathogenic *Escherichia coli*." International Journal of Medical Microbiology **295**(6-7): 471-478.

Le Gall, T., O. Clermont, S. Gouriou, B. Picard, X. Nassif, E. Denamur and O. Tenaillon (2007). "Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains." Molecular Biology and Evolution **24**(11): 2373-2384.

Lecointre, G., L. Rachdi, P. Darlu and E. Denamur (1998). "*Escherichia coli* molecular phylogeny using the incongruence length difference test." Molecular Biology and Evolution **15**(12): 1685-1695.

Lee, J. H., M. H. Cho and J. Lee (2011). "3-indolylacetonitrile decreases *Escherichia coli* O157:H7 biofilm formation and *Pseudomonas aeruginosa* virulence." Environ Microbiol **13**(1): 62-73.

Leopold, S. R., S. A. Sawyer, T. S. Whittam and P. I. Tarr (2011). "Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*." BMC Evol Biol **11**: 183.

Lerat, E. and H. Ochman (2004). "Psi-Phi: exploring the outer limits of bacterial pseudogenes." Genome Research **14**(11): 2273-2278.

Lescat, M., C. Hoede, O. Clermont, L. Garry, P. Darlu, P. Tuffery, E. Denamur and B. Picard (2009). "*aes*, the gene encoding the esterase B in *Escherichia coli*, is a powerful phylogenetic marker of the species." BMC Microbiology **9**: -.

Letunic, I. and P. Bork (2011). "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy." Nucleic Acids Res **39**(Web Server issue): W475-478.

Leveau, J. H. and S. E. Lindow (2001). "Appetite of an epiphyte: quantitative monitoring of bacterial sugar consumption in the phyllosphere." Proc Natl Acad Sci U S A **98**(6): 3446-3453.

Levert, M., O. Zamfir, O. Clermont, O. Bouvet, S. Lespinats, M. C. Hipeaux, C. Branger, B. Picard, C. Saint-Ruf, F. Norel, T. Balliau, M. Zivy, H. Le Nagard, S. Cruvellier, B. Chane-Woon-Ming, S. Nilsson, I. Gudelj, K. Phan, T. Ferenci, O. Tenaillon and E. Denamur (2010). "Molecular and Evolutionary Bases of Within-Patient Genotypic and Phenotypic Diversity in *Escherichia coli* Extraintestinal Infections." Plos Pathogens **6**(9): -.

Li, Y., R. E. Brackett, J. Chen and L. R. Beuchat (2001). "Survival and growth of *Escherichia coli* O157:H7 inoculated onto cut lettuce before or after heating in chlorinated water, followed by storage at 5 or 15 degrees C." J Food Prot **64**(3): 305-309.

Liao, C. H. (2008). "Growth of *Salmonella* on sprouting alfalfa seeds as affected by the inoculum size, native microbial load and *Pseudomonas fluorescens* 2-79." Lett Appl Microbiol **46**(2): 232-236.

Liao, C. H. and W. F. Fett (2001). "Analysis of native microflora and selection of strains antagonistic to human pathogens on fresh produce." J Food Prot **64**(8): 1110-1115.

Liao, D. (2000). "Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea." Journal of Molecular Evolution **51**(4): 305-317.

Lindow, S. E. and M. T. Brandl (2003). "Microbiology of the phyllosphere." Appl Environ Microbiol **69**(4): 1875-1883.

Lipsitch, M. and E. R. Moxon (1997). "Virulence and transmissibility of pathogens: what is the relationship?" Trends Microbiol **5**(1): 31-37.

Lloyd, A. L., D. A. Rasko and H. L. Mobley (2007). "Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*." Journal of Bacteriology **189**(9): 3532-3546.

Loper, J. E. and J. S. Buyer (1991). "Siderophores in Microbial Interactions on Plant Surfaces." Mol Plant Microbe Interact **4**(1): 5-13.

Loper, J. E. and M. D. Henkels (1999). "Utilization of heterologous siderophores enhances levels of iron available to *Pseudomonas putida* in the rhizosphere." Appl Environ Microbiol **65**(12): 5357-5363.

Lopez-Velasco, G., M. Davis, R. R. Boyer, R. C. Williams and M. A. Ponder (2010). "Alterations of the phylloepiphytic bacterial community associated with interactions of *Escherichia coli* O157:H7 during storage of packaged spinach at refrigeration temperatures." Food Microbiol **27**(4): 476-486.

Louws, F. J., D. W. Fulbright, C. T. Stephens and F. J. de Bruijn (1994). "Specific genomic fingerprints of phytopathogenic *Xanthomonas* and *Pseudomonas* pathovars and strains generated with repetitive sequences and PCR." Appl Environ Microbiol **60**(7): 2286-2295.

Lucchini, S., H. Liu, Q. Jin, J. C. Hinton and J. Yu (2005). "Transcriptional adaptation of *Shigella flexneri* during infection of macrophages and epithelial cells: insights into the strategies of a cytosolic bacterial pathogen." Infection and Immunity **73**(1): 88-102.

Lugtenberg, B. J., L. Dekkers and G. V. Bloemberg (2001). "Molecular determinants of rhizosphere colonization by *Pseudomonas*." Annu Rev Phytopathol **39**: 461-490.

Lukjancenko, O., T. M. Wassenaar and D. W. Ussery (2010). "Comparison of 61 sequenced *Escherichia coli* genomes." Microb Ecol **60**(4): 708-720.

Luo, C., S. T. Walk, D. M. Gordon, M. Feldgarden, J. M. Tiedje and K. T. Konstantinidis (2011). "Genome sequencing of environmental *Escherichia coli* expands

understanding of the ecology and speciation of the model bacterial species." Proc Natl Acad Sci U S A **108**(17): 7200-7205.

Luria, S. E. and J. W. Burrous (1957). "Hybridization between *Escherichia coli* and *Shigella*." Journal of Bacteriology **74**(4): 461-476.

Lynch, M. (1988). "Estimation of relatedness by DNA fingerprinting." Molecular Biology and Evolution **5**(5): 584-599.

Maharjan, R. P., S. Seeto and T. Ferenci (2007). "Divergence and redundancy of transport and metabolic rate-yield strategies in a single *Escherichia coli* population." Journal of Bacteriology **189**(6): 2350-2358.

Mandrell, R. E. (2009). Enteric human pathogens associated with fresh produce: sources, transport and ecology. Microbial safety of fresh produce: challenges, perspectives, and strategies. B. N. X. Fan, C. J. Doona, F. Feeherry, and R. B. Gravani (ed.), IFT/Blackwell Publishing, Ames, IA.**:** 3-42.

Martin, B., O. Humbert, M. Camara, E. Guenzi, J. Walker, T. Mitchell, P. Andrew, M. Prudhomme, G. Alloing, R. Hakenbeck and et al. (1992). "A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*." Nucleic Acids Res **20**(13): 3479-3483.

Martinez-Medina, M., X. Aldeguer, M. Lopez-Siles, F. Gonzalez-Huix, C. Lopez-Oliu, G. Dahbi, J. E. Blanco, J. Blanco, L. J. Garcia-Gil and A. Darfeuille-Michaud (2009). "Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease." Inflamm Bowel Dis **15**(6): 872-882.

Martins, M. T., I. G. Rivera, D. L. Clark, M. H. Stewart, R. L. Wolfe and B. H. Olson (1993). "Distribution of *uidA* gene sequences in *Escherichia coli* isolates in water sources and comparison with the expression of beta-glucuronidase activity in 4-methylumbelliferyl-beta-D-glucuronide media." Appl Environ Microbiol **59**(7): 2271-2276.

Matos, A., L. Kerkhof and J. L. Garland (2005). "Effects of microbial community diversity on the survival of *Pseudomonas aeruginosa* in the wheat rhizosphere." Microb Ecol **49**(2): 257-264.

Matte-Tailliez, O., C. Brochier, P. Forterre and H. Philippe (2002). "Archaeal phylogeny based on ribosomal proteins." Molecular Biology and Evolution **19**(5): 631-639.

Matthysse, A. G., R. Deora, M. Mishra and A. G. Torres (2008). "Polysaccharides cellulose, poly-beta-1,6-n-acetyl-D-glucosamine, and colanic acid are required for optimal binding of *Escherichia coli* O157:H7 strains to alfalfa sprouts and K-12 strains to plastic but not for binding to epithelial cells." Appl Environ Microbiol **74**(8): 2384-2390.

Matthysse, A. G., M. Marry, L. Krall, M. Kaye, B. E. Ramey, C. Fuqua and A. R. White (2005). "The effect of cellulose overproduction on binding and biofilm formation on roots by *Agrobacterium tumefaciens*." Mol Plant Microbe Interact **18**(9): 1002-1010.

Maurelli, A. T., R. E. Fernandez, C. A. Bloch, C. K. Rode and A. Fasano (1998). ""Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella spp*. and enteroinvasive *Escherichia coli*." Proc Natl Acad Sci U S A **95**(7): 3943-3948.

Maynard Smith, J., N. H. Smith, M. O'Rourke and B. G. Spratt (1993). "How clonal are bacteria?" Proc Natl Acad Sci U S A **90**(10): 4384-4388.

McDaniel, T. K., K. G. Jarvis, M. S. Donnenberg and J. B. Kaper (1995). "A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens." Proc Natl Acad Sci U S A **92**(5): 1664-1668.

Medini, D., C. Donati, H. Tettelin, V. Masignani and R. Rappuoli (2005). "The microbial pan-genome." Curr Opin Genet Dev **15**(6): 589-594.

Mellmann, A., D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg and H. Karch (2011). "Prospective genomic characterization of the German

enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology." PLoS One **6**(7): e22751.
Mercier, J. and S. E. Lindow (2000). "Role of leaf surface sugars in colonization of plants by bacterial epiphytes." Appl Environ Microbiol **66**(1): 369-374.
Messiha, N., A. van Diepeningen, N. Farag, S. Abdallah, J. Janse and A. van Bruggen (2007). "*Stenotrophomonas maltophilia*: a new potential biocontrol agent of *Ralstonia solanacearum*, causal agent of potato brown rot." European Journal of Plant Pathology **118**(3): 211-225.
Michino, H., K. Araki, S. Minami, S. Takaya, N. Sakai, M. Miyazaki, A. Ono and H. Yanagawa (1999). "Massive outbreak of *Escherichia coli* O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts." Am J Epidemiol **150**(8): 787-796.
Michod, R. E., H. Bernstein and A. M. Nedelcu (2008). "Adaptive value of sex in microbial pathogens." Infection Genetics and Evolution **8**(3): 267-285.
Millet, J., S. Baboolal, P. E. Akpaka, D. Ramoutar and N. Rastogi (2009). "Phylogeographical and molecular characterization of an emerging *Mycobacterium tuberculosis* clone in Trinidad and Tobago." Infection Genetics and Evolution **9**(6): 1336-1344.
Mira, A., H. Ochman and N. A. Moran (2001). "Deletional bias and the evolution of bacterial genomes." Trends Genet **17**(10): 589-596.
Miranda, R. L., T. Conway, M. P. Leatham, D. E. Chang, W. E. Norris, J. H. Allen, S. J. Stevenson, D. C. Laux and P. S. Cohen (2004). "Glycolytic and gluconeogenic growth of *Escherichia coli* O157:H7 (EDL933) and *E. coli* K-12 (MG1655) in the mouse intestine." Infection and Immunity **72**(3): 1666-1676.
Mitra, R., E. Cuesta-Alonso, A. Wayadande, J. Talley, S. Gilliland and J. Fletcher (2009). "Effect of route of introduction and host cultivar on the colonization, internalization, and movement of the human pathogen *Escherichia coli* O157:H7 in spinach." J Food Prot **72**(7): 1521-1530.
Monier, J. M. and S. E. Lindow (2004). "Frequency, size, and localization of bacterial aggregates on bean leaf surfaces." Appl Environ Microbiol **70**(1): 346-355.
Moran, N. A. (2002). "Microbial minimalism: genome reduction in bacterial pathogens." Cell **108**(5): 583-586.
Morgan, J. V. and H. B. Tukey (1964). "Characterization of Leachate from Plant Foliage." Plant Physiol **39**(4): 590-593.
Moritz, R. L. and R. A. Welch (2006). "The *Escherichia coli argW-dsdCXA* genetic island is highly variable, and *E. coli* K1 strains commonly possess two copies of *dsdCXA*." J Clin Microbiol **44**(11): 4038-4048.
Morris, C. E. and J. M. Monier (2003). "The ecological significance of biofilm formation by plant-associated bacteria." Annu Rev Phytopathol **41**: 429-453.
Mukherjee, A., M. K. Mammel, J. E. LeClerc and T. A. Cebula (2008). "Altered utilization of N-acetyl-D-galactosamine by *Escherichia coli* O157:H7 from the 2006 spinach outbreak." Journal of Bacteriology **190**(5): 1710-1717.
Mukherjee, A., D. Speh, E. Dyck and F. Diez-Gonzalez (2004). "Preharvest evaluation of coliforms, *Escherichia coli*, *Salmonella*, and *Escherichia coli* O157:H7 in organic and conventional produce grown by Minnesota farmers." J Food Prot **67**(5): 894-900.
Muyzer, G., E. C. de Waal and A. G. Uitterlinden (1993). "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA." Appl Environ Microbiol **59**(3): 695-700.
Nakabachi, A., A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran and M. Hattori (2006). "The 160-kilobase genome of the bacterial endosymbiont *Carsonella*." Science **314**(5797): 267.
Narra, H. P. and H. Ochman (2006). "Of what use is sex to bacteria?" Curr Biol **16**(17): R705-710.
Neefs, J. M., Y. Van de Peer, L. Hendriks and R. De Wachter (1990). "Compilation of small ribosomal subunit RNA sequences." Nucleic Acids Res **18 Suppl**: 2237-2317.

Niemira, B. A., C. J. Doona, F. E. Feeherry, X. Fan and R. B. Gravani (2009). Research Needs and Future Directions. Microbial Safety of Fresh Produce, Wiley-Blackwell**:** 419-425.

Nowrouzian, F. L., A. E. Wold and I. Adlerberth (2005). "*Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants." J Infect Dis **191**(7): 1078-1083.

O'Brien, A. D. and R. K. Holmes (1987). "Shiga and Shiga-like toxins." Microbiol Rev **51**(2): 206-220.

Ochman, H. and L. M. Davalos (2006). "The nature and dynamics of bacterial genomes." Science **311**(5768): 1730-1733.

Ochman, H. and R. K. Selander (1984). "Standard reference strains of *Escherichia coli* from natural populations." J Bacteriol **157**(2): 690-693.

Ochman, H., T. S. Whittam, D. A. Caugant and R. K. Selander (1983). "Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*." J Gen Microbiol **129**(9): 2715-2726.

Oliveira, M., J. Usall, I. Vinas, M. Anguera, F. Gatius and M. Abadias (2010). "Microbiological quality of fresh lettuce from organic and conventional production." Food Microbiol **27**(5): 679-684.

Olsen, A., A. Arnqvist, M. Hammar and S. Normark (1993). "Environmental regulation of curli production in *Escherichia coli*." Infect Agents Dis **2**(4): 272-274.

Palchevskiy, V. and S. E. Finkel (2006). "*Escherichia coli* competence gene homologs are essential for competitive fitness and the use of DNA as a nutrient." Journal of Bacteriology **188**(11): 3902-3910.

Payne, S. M. (1994). "Detection, isolation, and characterization of siderophores." Methods Enzymol **235**: 329-344.

Perez-Losada, M., M. L. Porter, L. Tazi and K. A. Crandall (2007). "New methods for inferring population dynamics from microbial sequences." Infection Genetics and Evolution **7**(1): 24-43.

Picard, B., J. S. Garcia, S. Gouriou, P. Duriez, N. Brahimi, E. Bingen, J. Elion and E. Denamur (1999). "The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection." Infection and Immunity **67**(2): 546-553.

Pinot, C., A. Deredjian, S. Nazaret, E. Brothier, B. Cournoyer, C. Segonds and S. Favre-Bonte (2011). "Identification of *Stenotrophomonas maltophilia* strains isolated from environmental and clinical samples: a rapid and efficient procedure." J Appl Microbiol.

Poulsen, L. K., F. Lan, C. S. Kristensen, P. Hobolth, S. Molin and K. A. Krogfelt (1994). "Spatial distribution of *Escherichia coli* in the mouse large intestine inferred from rRNA in situ hybridization." Infection and Immunity **62**(11): 5191-5194.

Poulsen, L. K., T. R. Licht, C. Rang, K. A. Krogfelt and S. Molin (1995). "Physiological state of *Escherichia coli* BJ4 growing in the large intestines of streptomycin-treated mice." Journal of Bacteriology **177**(20): 5840-5845.

Power, M. L., J. Littlefield-Wyer, D. M. Gordon, D. A. Veal and M. B. Slade (2005). "Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes." Environ Microbiol **7**(5): 631-640.

Preston-Mafham, J., L. Boddy and P. F. Randerson (2002). "Analysis of microbial community functional diversity using sole-carbon-source utilisation profiles - a critique." FEMS Microbiol Ecol **42**(1): 1-14.

Prigent-Combaret, C., G. Prensier, T. T. Le Thi, O. Vidal, P. Lejeune and C. Dorel (2000). "Developmental pathway for biofilm formation in curli-producing *Escherichia coli* strains: role of flagella, curli and colanic acid." Environ Microbiol **2**(4): 450-464.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J.

Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork and S. D. Ehrlich (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Quaiser, A., Y. Zivanovic, D. Moreira and P. Lopez-Garcia (2011). "Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara." ISME J **5**(2): 285-304.

Rai, P. K. and B. D. Tripathi (2007). "Microbial contamination in vegetables due to irrigation with partially treated municipal wastewater in a tropical city." Int J Environ Health Res **17**(5): 389-395.

Randazzo, C. L., G. O. Scifo, F. Tomaselli and C. Caggia (2009). "Polyphasic characterization of bacterial community in fresh cut salads." Int J Food Microbiol **128**(3): 484-490.

Rangel, J. M., P. H. Sparling, C. Crowe, P. M. Griffin and D. L. Swerdlow (2005). "Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982-2002." Emerging Infectious Diseases **11**(4): 603-609.

Rasko, D. A., M. J. Rosovitz, G. S. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio and J. Ravel (2008). "The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates." Journal of Bacteriology **190**(20): 6881-6893.

Rastogi, G., J. J. Tech, G. L. Coaker and J. H. Leveau (2010). "A PCR-based toolbox for the culture-independent quantification of total bacterial abundances in plant environments." J Microbiol Methods **83**(2): 127-132.

Redfield, R. J. (1993). "Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation." J Hered **84**(5): 400-404.

Redfield, R. J. (2001). "Do bacteria have sex?" Nat Rev Genet **2**(8): 634-639.

Rehfuss, M. Y., C. T. Parker and M. T. Brandl (2011). "*Salmonella* transcriptional signature in *Tetrahymena* phagosomes and role of acid tolerance in passage through the protist." ISME J **5**(2): 262-273.

Reid, S. J. and V. R. Abratt (2005). "Sucrose utilisation in bacteria: genetic organisation and regulation." Appl Microbiol Biotechnol **67**(3): 312-321.

Retchless, A. C. and J. G. Lawrence (2010). "Phylogenetic incongruence arising from fragmented speciation in enteric bacteria." Proc Natl Acad Sci U S A **107**(25): 11453-11458.

Rice, E. W., C. H. Johnson, D. K. Wild and D. J. Reasoner (1992). "Survival of *Escherichia coli* O157: H7 in drinking water associated with a waterborne disease outbreak of hemorrhagic colitis." Letters in Applied Microbiology **15**(2): 38-40.

Rodger, G. and J. P. Blakeman (1984). "Microbial colonization and uptake of 14C label on leaves of sycamore." Transactions of the British Mycological Society **82**(1): 45-51.

Rohde, H., J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F. Gao, Y. Song, H. Yang, J. Wang, J. Xu, M. J. Pallen, M. Aepfelbacher and R. Yang (2011). "Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4." N Engl J Med **365**(8): 718-724.

Roos, V., G. C. Ulett, M. A. Schembri and P. Klemm (2006). "The asymptomatic bacteriuria *Escherichia coli* strain 83972 outcompetes uropathogenic *E. coli* strains in human urine." Infection and Immunity **74**(1): 615-624.

Russo, T. A. and J. R. Johnson (2003). "Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem." Microbes and Infection **5**(5): 449-456.

Ryan, R. P., S. Monchy, M. Cardinale, S. Taghavi, L. Crossman, M. B. Avison, G. Berg, D. van der Lelie and J. M. Dow (2009). "The versatility and adaptation of bacteria from the genus *Stenotrophomonas*." Nature Reviews Microbiology **7**(7): 514-525.

Sabarly, V., O. Bouvet, J. Glodt, O. Clermont, D. Skurnik, L. Diancourt, D. De Vienne, E. Denamur and C. Dillmann (2011). "The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity." J Evol Biol **24**(7): 1559-1571.

Sagoo, S. K., C. L. Little and R. T. Mitchell (2001). "The microbiological examination of ready-to-eat organic vegetables from retail establishments in the United Kingdom." Lett Appl Microbiol **33**(6): 434-439.

Sankar, T. S., G. Neelakanta, V. Sangal, G. Plum, M. Achtman and K. Schnetz (2009). "Fate of the H-NS-repressed *bgl* operon in evolution of *Escherichia coli*." Plos Genetics **5**(3): e1000405.

Sasaki, T., M. Kobayashi and N. Agui (2000). "Epidemiological potential of excretion and regurgitation by *Musca domestica* (Diptera: Muscidae) in the dissemination of *Escherichia coli* O157: H7 to food." J Med Entomol **37**(6): 945-949.

Savageau, M. A. (1974). "Genetic regulatory mechanisms and the ecological niche of *Escherichia coli*." Proc Natl Acad Sci U S A **71**(6): 2453-2455.

Savageau, M. A. (1983). "*Escherichia coli* Habitats, Cell Types, and Molecular Mechanisms of Gene-Control." Am Nat **122**(6): 732-744.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schikora, A., A. Carreri, E. Charpentier and H. Hirt (2008). "The dark side of the salad: *Salmonella typhimurium* overcomes the innate immune response of *Arabidopsis thaliana* and shows an endopathogenic lifestyle." PLoS One **3**(5): e2279.

Schultz, M. (2008). "Clinical use of *E. coli* Nissle 1917 in inflammatory bowel disease." Inflamm Bowel Dis **14**(7): 1012-1018.

Scouten, A. J. and L. R. Beuchat (2002). "Combined effects of chemical, heat and ultrasound treatments to kill *Salmonella* and *Escherichia coli* O157:H7 on alfalfa seeds." J Appl Microbiol **92**(4): 668-674.

Sears, H. J. and I. Brownlee (1952). "Further observations on the persistence of individual strains of *Escherichia coli* in the intestinal tract of man." Journal of Bacteriology **63**(1): 47-57.

Sears, H. J., I. Brownlee and J. K. Uchiyama (1950). "Persistence of individual strains of *Escherichia coli* in the intestinal tract of man." Journal of Bacteriology **59**(2): 293-301.

Sela, S., D. Nestel, R. Pinto, E. Nemny-Lavy and M. Bar-Joseph (2005). "Mediterranean fruit fly as a potential vector of bacterial pathogens." Appl Environ Microbiol **71**(7): 4052-4056.

Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour and T. S. Whittam (1986). "Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics." Appl Environ Microbiol **51**(5): 873-884.

Selander, R. K. and B. R. Levin (1980). "Genetic diversity and structure in *Escherichia coli* populations." Science **210**(4469): 545-547.

Serres, M. H., A. R. Kerr, T. J. McCormack and M. Riley (2009). "Evolution by leaps: gene duplication in bacteria." Biol Direct **4**: 46.

Shaw, R. K., C. N. Berger, B. Feys, S. Knutton, M. J. Pallen and G. Frankel (2008). "Enterohemorrhagic *Escherichia coli* exploits EspA filaments for attachment to salad leaves." Appl Environ Microbiol **74**(9): 2908-2914.

Shaw, R. K., C. N. Berger, M. J. Pallen, Å. Sjöling and G. Frankel (2011). "Flagella mediate attachment of enterotoxigenic *Escherichia coli* to fresh salad leaves." Env Microbiol Rep **3**(1): 112-117.

Sheldon, I. M., A. N. Rycroft, B. Dogan, M. Craven, J. J. Bromfield, A. Chandler, M. H. Roberts, S. B. Price, R. O. Gilbert and K. W. Simpson (2010). "Specific strains of *Escherichia coli* are pathogenic for the endometrium of cattle and cause pelvic inflammatory disease in cattle and mice." PLoS One **5**(2): e9192.

Shen, P. and H. V. Huang (1986). "Homologous recombination in *Escherichia coli*: dependence on substrate length and homology." Genetics **112**(3): 441-457.
Shen, Z., W. Qu, W. Wang, Y. Lu, Y. Wu, Z. Li, X. Hang, X. Wang, D. Zhao and C. Zhang (2010). "MPprimer: a program for reliable multiplex PCR primer design." BMC Bioinformatics **11**: 143.
Shulman, S. T., H. C. Friedmann and R. H. Sims (2007). "Theodor Escherich: the first pediatric infectious diseases physician?" Clinical Infectious Diseases **45**(8): 1025-1029.
Sims, G. E. and S. H. Kim (2011). "Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs)." Proc Natl Acad Sci U S A **108**(20): 8329-8334.
Sjogren, R. E. (1995). "Thirteen-year survival study of an environmental *Escherichia coli* in field mini-plots." Water, Air, & Soil Pollution **81**(3): 315-335.
Skippington, E. and M. A. Ragan (2011). "Lateral genetic transfer and the construction of genetic exchange communities." FEMS Microbiol Rev **35**(5): 707-735.
Smith, M. G. (1975). "In vivo transfer of R factors between *Escherichia coli* strains inoculated into the rumen of sheep." J Hyg (Lond) **75**(3): 363-370.
Smith, S. N., E. C. Hagan, M. C. Lane and H. L. Mobley (2010). "Dissemination and systemic colonization of uropathogenic *Escherichia coli* in a murine model of bacteremia." MBio **1**(5).
Soderstrom, A., A. Lindberg and Y. Andersson (2005). "EHEC O157 outbreak in Sweden from locally produced lettuce, August-September 2005." Euro Surveill **10**(9): E050922 050921.
Solo-Gabriele, H. M., M. A. Wolfert, T. R. Desmarais and C. J. Palmer (2000). "Sources of *Escherichia coli* in a coastal subtropical environment." Appl Environ Microbiol **66**(1): 230-237.
Solomon, E. B., H. J. Pang and K. R. Matthews (2003). "Persistence of *Escherichia coli* O157:H7 on lettuce plants following spray irrigation with contaminated water." J Food Prot **66**(12): 2198-2202.
Spratt, B. G. (2004). "Exploring the concept of clonality in bacteria." Methods Mol Biol **266**: 323-352.
Stabler, R., L. Dawson and B. Wren (2010). "Comparative Genome Analysis of *Clostridium difficile* using DNA microarrays." Methods Mol Biol **646**: 149-162.
Staley, T. E., E. W. Jones and L. D. Corley (1969). "Attachment and penetration of *Escherichia coli* into intestinal epithelium of the ileum in newborn pigs." Am J Pathol **56**(3): 371-392.
Steele, M. and J. Odumeru (2004). "Irrigation water as source of foodborne pathogens on fruit and vegetables." J Food Prot **67**(12): 2839-2849.
Stentz, R., U. Wegmann, M. Parker, R. Bongaerts, L. Lesaint, M. Gasson and C. Shearman (2009). "CsiA is a bacterial cell wall synthesis inhibitor contributing to DNA translocation through the cell envelope." Mol Microbiol **72**(3): 779-794.
Suckstorff, I. and G. Berg (2003). "Evidence for dose-dependent effects on plant growth by *Stenotrophomonas* strains from different origins." J Appl Microbiol **95**(4): 656-663.
Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran and S. G. Andersson (2002). "50 million years of genomic stasis in endosymbiotic bacteria." Science **296**(5577): 2376-2379.
Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar (2011). "MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." Mol Biol Evol **In press**.
Taormina, P. J. and L. R. Beuchat (1999). "Comparison of chemical treatments to eliminate enterohemorrhagic *Escherichia coli* O157:H7 on alfalfa seeds." J Food Prot **62**(4): 318-324.
Tenaillon, O., D. Skurnik, B. Picard and E. Denamur (2010). "The population genetics of commensal *Escherichia coli*." Nat Rev Microbiol **8**(3): 207-217.

Teplitski, M., J. D. Barak and K. R. Schneider (2009). "Human enteric pathogens in produce: un-answered ecological questions with direct implications for food safety." Curr Opin Biotechnol **20**(2): 166-171.

Texier, S., C. Prigent-Combaret, M. H. Gourdon, M. A. Poirier, P. Faivre, J. M. Dorioz, J. Poulenard, L. Jocteur-Monrozier, Y. Moenne-Loccoz and D. Trevisan (2008). "Persistence of culturable *Escherichia coli* fecal contaminants in dairy alpine grassland soils." J Environ Qual **37**(6): 2299-2310.

Thomas, C. M. and K. M. Nielsen (2005). "Mechanisms of, and barriers to, horizontal gene transfer between bacteria." Nature Reviews Microbiology **3**(9): 711-721.

Thompson, A., S. Lucchini and J. C. Hinton (2001). "It's easy to build your own microarrayer!" Trends Microbiol **9**(4): 154-156.

Tindall, K. R. and T. A. Kunkel (1988). "Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase." Biochemistry **27**(16): 6008-6013.

Tobes, R. and J. L. Ramos (2005). "REP code: defining bacterial identity in extragenic space." Environ Microbiol **7**(2): 225-228.

Torres, A. G., C. Jeter, W. Langley and A. G. Matthysse (2005). "Differential binding of *Escherichia coli* O157:H7 to alfalfa, human epithelial cells, and plastic is mediated by a variety of surface structures." Appl Environ Microbiol **71**(12): 8008-8015.

Torriani, S., C. Orsi and M. Vescovo (1997). "Potential of *Lactobacillus casei*, Culture Permeate, and Lactic Acid To Control Microorganisms in Ready-To-Use Vegetables." Journal of Food Protection **60**(12): 1564-1567.

Toth, I. K., L. Pritchard and P. R. Birch (2006). "Comparative genomics reveals what makes an enterobacterial plant pathogen." Annu Rev Phytopathol **44**: 305-336.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. El Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat, S. Mangenot, V. Martinez-Jehanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. Saint Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Medigue, E. P. C. Rocha and E. Denamur (2009). "Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths." Plos Genetics **5**(1): -.

Tourlomousis, P., E. K. Kemsley, K. P. Ridgway, M. J. Toscano, T. J. Humphrey and A. Narbad (2010). "PCR-denaturing gradient gel electrophoresis of complex microbial communities: a two-step approach to address the effect of gel-to-gel variation and allow valid comparisons across a large dataset." Microb Ecol **59**(4): 776-786.

Tullus, K., I. Kuhn, I. Orskov, F. Orskov and R. Mollby (1992). "The importance of P and type 1 fimbriae for the persistence of *Escherichia coli* in the human gut." Epidemiol Infect **108**(3): 415-421.

Tyrrel, S. F., J. W. Knox and E. K. Weatherhead (2006). "Microbiological Water Quality Requirements for Salad Irrigation in the United Kingdom." Journal of Food Protection **69**(8): 2029-2035.

Uhlich, G. A., J. E. Keen and R. O. Elder (2001). "Mutations in the *csgD* promoter associated with variations in curli expression in certain strains of *Escherichia coli* O157:H7." Appl Environ Microbiol **67**(5): 2367-2370.

Uhlich, G. A., J. R. Sinclair, N. G. Warren, W. A. Chmielecki and P. Fratamico (2008). "Characterization of Shiga toxin-producing *Escherichia coli* isolates associated with two multistate food-borne outbreaks that occurred in 2006." Appl Environ Microbiol **74**(4): 1268-1272.

Ukena, S. N., A. Singh, U. Dringenberg, R. Engelhardt, U. Seidler, W. Hansen, A. Bleich, D. Bruder, A. Franzke, G. Rogler, S. Suerbaum, J. Buer, F. Gunzer and A. M. Westendorf (2007). "Probiotic *Escherichia coli* Nissle 1917 inhibits leaky gut by enhancing mucosal integrity." PLoS One **2**(12): e1308.

Ussery, D. W., T. T. Binnewies, R. Gouveia-Oliveira, H. Jarmer and P. F. Hallin (2004). "Genome update: DNA repeats in bacterial genomes." Microbiology **150**(Pt 11): 3519-3521.

Valdebenito, M., A. L. Crumbliss, G. Winkelmann and K. Hantke (2006). "Environmental factors influence the production of enterobactin, salmochelin, aerobactin, and yersiniabactin in *Escherichia coli* strain Nissle 1917." International Journal of Medical Microbiology **296**(8): 513-520.

Valentin-Bon, I., A. Jacobson, S. R. Monday and P. C. Feng (2008). "Microbiological quality of bagged cut spinach and lettuce mixes." Appl Environ Microbiol **74**(4): 1240-1242.

van Belkum, A., M. Sluijuter, R. de Groot, H. Verbrugh and P. W. Hermans (1996). "Novel BOX repeat PCR assay for high-resolution typing of *Streptococcus pneumoniae* strains." Journal of Clinical Microbiology **34**(5): 1176-1179.

van Elsas, J. D., A. V. Semenov, R. Costa and J. T. Trevors (2011). "Survival of *Escherichia coli* in the environment: fundamental and public health aspects." ISME J **5**(2): 173-183.

van Passel, M. W., P. R. Marri and H. Ochman (2008). "The emergence and fate of horizontally acquired genes in *Escherichia coli*." PLoS Comput Biol **4**(4): e1000059.

Versalovic, J., C. R. Woods, Jr., P. R. Georghiou, R. J. Hamill and J. R. Lupski (1993). "DNA-based identification and epidemiologic typing of bacterial pathogens." Arch Pathol Lab Med **117**(11): 1088-1098.

Vescovo, M., C. Orsi, G. Scolari and S. Torriani (1995). "Inhibitory effect of selected lactic acid bacteria on microflora associated with ready-to-use vegetables." Lett Appl Microbiol **21**(2): 121-125.

Vescovo, M., S. Torriani, C. Orsi, F. Macchiarolo and G. Scolari (1996). "Application of antimicrobial-producing lactic acid bacteria to control pathogens in ready-to-use vegetables." J Appl Bacteriol **81**(2): 113-119.

Vidal, O., R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman and P. Lejeune (1998). "Isolation of an *Escherichia coli* K-12 mutant strain able to form biofilms on inert surfaces: involvement of a new *ompR* allele that increases curli expression." Journal of Bacteriology **180**(9): 2442-2449.

Vital, M., F. Hammes and T. Egli (2008). "*Escherichia coli* O157 can grow in natural freshwater at low carbon concentrations." Environ Microbiol **10**(9): 2387-2396.

Vogel-Scheel, J., C. Alpert, W. Engst, G. Loh and M. Blaut (2010). "Requirement of purine and pyrimidine synthesis for colonization of the mouse intestine by *Escherichia coli*." Appl Environ Microbiol **76**(15): 5181-5187.

Vos, M. and X. Didelot (2009). "A comparison of homologous recombination rates in bacteria and archaea." ISME J **3**(2): 199-208.

Vyas, P., R. Joshi, K. C. Sharma, P. Rahi and A. Gulati (2010). "Cold-adapted and rhizosphere-competent strain of *Rahnella sp.* with broad-spectrum plant growth-promotion potential." J Microbiol Biotechnol **20**(12): 1724-1734.

Wachtel, M. R., L. C. Whitehand and R. E. Mandrell (2002). "Association of *Escherichia coli* O157:H7 with preharvest leaf lettuce upon exposure to contaminated irrigation water." J Food Prot **65**(1): 18-25.

Walk, S. T., E. W. Alm, L. M. Calhoun, J. M. Mladonicky and T. S. Whittam (2007). "Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches." Environ Microbiol **9**(9): 2274-2288.

Walk, S. T., E. W. Alm, D. M. Gordon, J. L. Ram, G. A. Toranzos, J. M. Tiedje and T. S. Whittam (2009). "Cryptic lineages of the genus *Escherichia*." Appl Environ Microbiol **75**(20): 6534-6544.

Walk, S. T., J. M. Mladonicky, J. A. Middleton, A. J. Heidt, J. R. Cunningham, P. Bartlett, K. Sato and T. S. Whittam (2007). "Influence of antibiotic selection on genetic composition of *Escherichia coli* populations from conventional and organic dairy farms." Appl Environ Microbiol **73**(19): 5982-5989.

Wallick, H. and C. A. Stuart (1943). "Antigenic Relationships of *Escherichia coli* Isolated from One Individual." Journal of Bacteriology **45**(2): 121-126.

Warriner, K., F. Ibrahim, M. Dickinson, C. Wright and W. M. Waites (2003). "Interaction of *Escherichia coli* with growing salad spinach plants." J Food Prot **66**(10): 1790-1797.

Warriner, K., F. Ibrahim, M. Dickinson, C. Wright and W. M. Waites (2003). "Internalization of human pathogens within growing salad vegetables." Biotechnol Genet Eng Rev **20**: 117-134.

Warriner, K. and A. Namvar (2010). "The tricks learnt by human enteric pathogens from phytopathogens to persist within the plant environment." Curr Opin Biotechnol **21**(2): 131-136.

Waterman, S. R. and P. L. Small (1996). "Characterization of the acid resistance phenotype and *rpoS* alleles of shiga-like toxin-producing *Escherichia coli*." Infect Immun **64**(7): 2808-2811.

Werren, J. H. (2011). "Selfish genetic elements, genetic conflict, and evolutionary innovation." Proc Natl Acad Sci U S A **108 Suppl 2**: 10863-10870.

Whipps, J. M., P. Hand, D. A. Pink and G. D. Bending (2008). "Human pathogens and the phyllosphere." Adv Appl Microbiol **64**: 183-221.

White, A. P., D. L. Gibson, G. A. Grassl, W. W. Kay, B. B. Finlay, B. A. Vallance and M. G. Surette (2008). "Aggregation via the red, dry, and rough morphotype is not a virulence adaptation in *Salmonella enterica* serovar Typhimurium." Infection and Immunity **76**(3): 1048-1058.

White, A. P., D. L. Gibson, W. Kim, W. W. Kay and M. G. Surette (2006). "Thin aggregative fimbriae and cellulose enhance long-term survival and persistence of *Salmonella*." Journal of Bacteriology **188**(9): 3219-3227.

White, A. P., K. A. Sibley, C. D. Sibley, J. D. Wasmuth, R. Schaefer, M. G. Surette, T. A. Edge and N. F. Neumann (2011). "Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity." Appl Environ Microbiol.

Whitman, R. L. and M. B. Nevers (2003). "Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach." Appl Environ Microbiol **69**(9): 5555-5562.

Whitman, R. L., D. A. Shively, H. Pawlik, M. B. Nevers and M. N. Byappanahalli (2003). "Occurrence of *Escherichia coli* and enterococci in Cladophora (Chlorophyta) in nearshore water and beach sand of Lake Michigan." Appl Environ Microbiol **69**(8): 4714-4719.

Whittam, T. S. (1989). "Clonal dynamics of *Escherichia coli* in its natural habitat." Antonie Van Leeuwenhoek **55**(1): 23-32.

Whittam, T. S., H. Ochman and R. K. Selander (1983). "Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*." Molecular Biology and Evolution **1**(1): 67-83.

Whittam, T. S., H. Ochman and R. K. Selander (1983). "Multilocus genetic structure in natural populations of *Escherichia coli*." Proc Natl Acad Sci U S A **80**(6): 1751-1755.

Wick, L. M., W. Qi, D. W. Lacher and T. S. Whittam (2005). "Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7." Journal of Bacteriology **187**(5): 1783-1791.

Wiles, T. J., R. R. Kulesus and M. A. Mulvey (2008). "Origins and virulence mechanisms of uropathogenic *Escherichia coli*." Exp Mol Pathol **85**(1): 11-19.

Wilson, M. and S. E. Lindow (1994). "Coexistence among Epiphytic Bacterial Populations Mediated through Nutritional Resource Partitioning." Appl Environ Microbiol **60**(12): 4468-4477.

Wilson, M., M. A. Savka, I. Hwang, S. K. Farrand and S. E. Lindow (1995). "Altered Epiphytic Colonization of Mannityl Opine-Producing Transgenic Tobacco Plants by a Mannityl Opine-Catabolizing Strain of *Pseudomonas syringae*." Appl Environ Microbiol **61**(6): 2151-2158.

Winfield, M. D. and E. A. Groisman (2003). "Role of nonhost environments in the lifestyles of *Salmonella* and *Escherichia coli*." Appl Environ Microbiol **69**(7): 3687-3694.

Wirth, T., D. Falush, R. Lan, F. Colles, P. Mensa, L. H. Wieler, H. Karch, P. R. Reeves, M. C. Maiden, H. Ochman and M. Achtman (2006). "Sex and virulence in *Escherichia coli*: an evolutionary perspective." Mol Microbiol **60**(5): 1136-1151.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proc Natl Acad Sci U S A **74**(11): 5088-5090.

Wold, A. E., D. A. Caugant, G. Lidin-Janson, P. de Man and C. Svanborg (1992). "Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics." Journal of Infectious Diseases **165**(1): 46-52.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk and J. A. Eisen (2009). "A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*." Nature **462**(7276): 1056-1060.

Xicohtencatl-Cortes, J., E. Sanchez Chacon, Z. Saldana, E. Freer and J. A. Giron (2009). "Interaction of *Escherichia coli* O157:H7 with leafy green produce." J Food Prot **72**(7): 1531-1537.

Yanai, I., C. J. Camacho and C. DeLisi (2000). "Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification." Phys Rev Lett **85**(12): 2641-2644.

Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." Trends Ecol Evol **15**(12): 496-503.

Yu, Z. and M. Morrison (2004). "Comparisons of different hypervariable regions of rrs genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis." Appl Environ Microbiol **70**(8): 4800-4806.

Zdziarski, J., C. Svanborg, B. Wullt, J. Hacker and U. Dobrindt (2008). "Molecular basis of commensalism in the urinary tract: low virulence or virulence attenuation?" Infection and Immunity **76**(2): 695-703.

Zhang, L., B. Foxman and C. Marrs (2002). "Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group B2." J Clin Microbiol **40**(11): 3951-3955.

Zipfel, C. (2008). "Pattern-recognition receptors in plant innate immunity." Curr Opin Immunol **20**(1): 10-16.