Ecological and genetic determinants of plasmid distribution in Escherichia coli.

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Environmental Microbiology and Environmental Microbiology Reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>EMI-2016-0603.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>EMI - Research article</td>
</tr>
<tr>
<td>Journal:</td>
<td>Environmental Microbiology</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Medaney, Frances; Royal Holloway University of London School of Biological Sciences Ellis, Richard; Animal and Plant Health Agency Raymond, Ben; University of Exeter, College of Life and Environmental Sciences</td>
</tr>
<tr>
<td>Keywords:</td>
<td>antibiotic resistance, mobile elements, population structure, plasmid profile, replicon typing</td>
</tr>
</tbody>
</table>
Ecological and genetic determinants of plasmid distribution in *Escherichia coli*.

Frances Medaney¹, Richard J. Ellis², Ben Raymond¹³*.

¹School of Biological Science, Royal Holloway University of London, Egham, Surrey, TW20 0EX, UK.  ²Specialist Scientific Support Department, Animal and Plant Health Agency, APHA Weybridge, Addlestone, Surrey, KT15 3NB, UK. ³University of Exeter, Penrun campus, Penryn, Cornwall, TR10 9FE, UK. *corresponding author, b.raymond@exeter.ac.uk; tel +441326259085.

Running title:

The fundamental ecology of *E. coli* plasmids
Originality Significance statement:

Large conjugative plasmids are clearly very important as carriers of virulence and antimicrobial resistance genes. Nevertheless the vast majority of research on plasmids has focussed on genetics, while study of plasmid ecology and evolutionary biology is in its infancy. Plasmid ecology and evolution are fundamentally important as they can explain how and why plasmids move through bacterial communities and why they might be retained or lost by particular lineages.

Theoretical and experimental evolution studies on plasmid evolutionary ecology are relatively common but we have little basis for knowing whether the ecological assumptions these studies make are realistic. The motivation for this study was the lack data on the fundamental field ecology of large plasmids. Previous studies of plasmid “ecology” have typically been either fishing for novel resistance plasmids, or surveying plasmids in populations recently exposed to antibiotics, are under-replicated or take no account of bacterial population structure. High throughput next generation sequencing approaches have also not been particularly helpful as they tend to loose vital data on bacteria plasmid associations.

The fundamental field ecology approach we have taken is original as it incorporates a fully replicated hierarchical design that characterizes plasmid profiles within a bacterial population genetic structure; this allows us to make inferences about how both ecological and genetic factors can shape plasmid profiles. Novel findings include the effect of bacterial diversity on the diversity of plasmids found in a community; the fact that different plasmid replicons have different host ranges within species; the fact that the presence of large replicons can increase or decrease the likelihood of carriage of other replicons; the effect of site on the probability of carriage and the substantial and unexplained effect of bacterial genetic background on the
probability that isolates will have no large plasmids. As an original basic study of
plasmid field ecology this work generates a large number of new hypotheses about how
plasmids and bacteria might interact with each other

Summary

Bacterial plasmids are important carriers of virulence and antibiotic resistance genes.
Nevertheless, we know little of the determinants of plasmid distribution in bacterial
populations. Here we explored the factors affecting the diversity and distribution of the
large plasmids of *Escherichia coli* in populations of cattle grazing on semi-natural
grassland, a set of populations with low frequencies of antibiotic resistance genes.

Critically, we also characterized the population genetic structure of bacterial hosts. This
revealed structured *E. coli* populations with high diversity between sites and individuals
but low diversity within cattle hosts. Plasmid profiles, however, varied considerably
within the same *E. coli* genotype. Both ecological and genetic factors affected plasmid
distribution: plasmid profiles were affected by site, *E. coli* diversity, *E. coli* genotype and
the presence of other large plasmids. Notably 3/26 *E. coli* serotypes accounted for half
the observed plasmid-free isolates indicating that within species variation can
substantially affect carriage of the major conjugative plasmids. The observed population
structure suggest that most of the opportunities for within species plasmid transfer
occur between different individuals of the same genotype and support recent
experimental work indicating that plasmid-host coevolution, and epistatic interactions
on fitness costs are likely to be important in determining occupancy.

Keywords: antibiotic resistance, mobile elements, population structure, plasmid profile,
replicon typing
Introduction

The persistence, replication and spread of bacterial plasmids is a major focus of research because so many plasmids carry virulence and antibiotic resistance genes (Carattoli, 2011; Dröge et al., 2000; Hopkins et al., 2006; Johnson et al., 2007; Marcadé et al., 2009; Pallecchi et al., 2007). Despite their clinical importance, the ecology and evolutionary biology of plasmids has received relatively little attention. There are a range of contrasting views of the relationship between plasmids and their bacterial hosts. We can think of plasmids as genetic parasites, costly junk DNA that offset the costs they impose on their hosts by conjugating and infecting new hosts (Lili et al., 2007; Modi and Adams, 1991; Smith, 2010). Plasmids can also be seen as carriers of intermittently selected traits (Eberhard, 1990; Lilley and Bailey, 1997b), or indeed as "domesticated" accessory genomes that may end up loosing conjugational ability entirely, such as the plasmid pXO2 that encodes encapsulation genes for Bacillus anthracis (Hu et al., 2009).

Understanding how plasmids can persist in bacterial populations over the long term is an evolutionary dilemma referred to as the "plasmid paradox". The persistence of plasmids and their association with periodically beneficial traits is a complex and developing theoretical area. Theory suggests that if plasmids are on average beneficial to their hosts, because of the traits they carry, then selection should favour bacteria that transfer these beneficial traits to the chromosome, as they can then avoid the costs of plasmid carriage (Bergstrom et al., 2000). Conversely lack of selection for non-essential plasmid encoded traits can lead to their loss from the plasmid itself (Dahlberg and Chao, 2003). In the view of some, the costs incurred by plasmids, and segregational loss during vertical transfer, means that it is difficult for plasmids that do not encode beneficial traits to persist (Bergstrom et al., 2000).
Emerging evidence is helping to elucidate why plasmids are common and tend to be associated with particular classes of traits. Studies of conjugation rates in biofilms, diverse communities or semi-natural settings have shown that these rates can be much higher than previously thought (Dionisio et al., 2002; Luo et al., 2005; Sorensen et al., 2005; Vilas-Bôas et al., 2008). In addition, the ability of bacterial chromosomes to rapidly adapt and ameliorate or negate the fitness costs of plasmid carriage seems to be a common pattern in ensuring plasmid persistence (Dahlberg and Chao, 2003; Dionisio et al., 2005; Harrison et al., 2015; San Millan et al., 2014). While these varied viewpoints may reflect real variation in bacterial plasmid associations it is worthwhile asking how applicable these views are to natural populations of bacteria and hosts. For example, what proportion of plasmids carry intermittently selected traits? Are bacterial lineages associating with plasmids for long enough to lead to substantial selection pressure on costs of carriage? If bacteria can adapt to ameliorate the fitness costs of plasmid carriage what then limits the prevalence and host range of plasmids? Exploring their fundamental ecology can help provide basic information for evaluating current theory but also help us develop novel theory to explain the variation that we see.

Our current understanding of plasmid ecology has been biased by the common means of collection: most plasmids are collected using selective agents that isolate bacteria carrying particular resistance genes, typically for antibiotics, heavy metals or other pollutants (Dierikx et al., 2012; Lilley and Bailey, 1997a). Plasmid multi locus sequence typing (pMLST) has been used to identify the prevalence of resistance plasmids in livestock, and has demonstrated the presence of identical extended spectrum beta-lactamase (ESBL) genes and plasmids in livestock and humans, suggesting the transfer of bacteria and plasmids through the food chain (Dierikx et al., 2010; Dierikx et al., 2012; Hordijk et al., 2013; Leverstein-van Hall et al., 2011).
Another common approach is to recover plasmids from environment that may recently have been subject to strong selection for antibiotic resistance (Dierikx et al., 2010; Dierikx et al., 2012; Dröge et al., 2000). However, these studies are not usually designed to assess bacterial population structure and so sample sizes can be small, and herd and host animal replication low, (Dierikx et al., 2010; Hordijk et al., 2013). In addition, bacterial populations may have low diversity if they have recently been subject to strong selective sweeps, and we are left knowing little about the distribution and ecology of plasmids that do not carry these intermittently selected traits. Exceptions, i.e. plasmid studies not focussed on resistance genes, tend to be from the older literature, when studies were limited to low replication and examined very broad species / plasmid associations (Datta and Hughes, 1983) or are modern deep-sequencing approaches targeting small circular molecules, in which bacterial host-association information is typically lost (Kav et al., 2012).

Thus, despite years of study of plasmids and their genetics we have few data on how bacterial hosts and ecological factors determine distribution. Importantly, this may mean that we do not know what will happen to plasmids after selection for resistance has ceased. The aims of this study were to explore the fundamental ecology of *Escherichia coli* plasmids in a natural context, without strong selection, and to explore how variation in bacterial host populations, particularly genetic diversity and abundance, could affect the diversity and distribution of plasmid carriage. We selected a study system based on isolating culturable strains from fresh bovine faecal matter produced by cattle grazing in unimproved semi-natural grassland, on either common land or land managed for conservation. We anticipated that these outdoor reared animals would maintain diverse enteric populations that had not been subject to strong selection for antibiotic resistance. This study system permitted a hierarchical sampling...
strategy (plasmids within bacteria within pats within fields) that could be replicated
across multiple sites with independent cattle populations.

4 Results

Diversity and antibiotic resistance in the *E. coli* populations.

We characterized the bacterial population in order to explore levels of antibiotic
resistance; to confirm that our bacteria were comprised of distinct populations and to
assess the distribution of bacterial diversity between sites and vertebrate hosts. In total
24483 colonies were isolated from 84 cowpats across 9 sites. Of these, 527 isolates were
genotyped and screened for plasmid replicons and antibiotic susceptibility. Populations
were diverse, strongly structured and characterized by a low frequency of resistance
isolates. Of the 527 isolates just 14 (less than 3%) were resistant to any of the
antibiotics tested: 9 isolates were ampicillin resistant, 4 were chloramphenicol resistant,
and just 1 was resistant to cefotaxime. No multi-resistant isolates were found. Resistant
isolates were found in 5 of the 9 sites, with four sites resistance-free.

Bacterial populations were highly aggregated with the majority of the population
concentrated in relatively few samples; just 7/84 pats contributed more than 50% of the
total colony counts. Within each pat, there was relatively low within species diversity,
the vast majority of pats contained one or two H-antigen genotypes (Fig 1A). However,
there was opportunity for plasmid transfer within genotypes within bovine hosts. We
examined the range in numbers of replicons between isolates for pats that had at least 4
isolates of the same serotype. Most commonly plasmid carriage was not at fixation, in
37 pats there was heterogeneous plasmid carriage within bacterial genotypes, while
only 20 pats showed uniform plasmid profiles (Fig 1B).
Different sites contained distinctive aggregations of *E. coli* genotypes (Fig 1C) confirming that these were good biological replicates. Hierarchical $F$-statistics were calculated to assess the contribution of sites, and pats within sites to the *E. coli* population structure: Both pat within site ($F_{\text{pat/site}} = 0.457, p < 1 \times 10^{-4}$), and site ($F_{\text{site/total}} = 0.197, p < 1 \times 10^{-4}$) were found to have strong effects on the population structure of *E. coli* (10,000 permutations). In other words, 20% of the observed genotype variation can be explained by site, and 46% of this variation can be explained between pats within site, corresponding to great and very great genetic differentiation respectively (Holsinger and Weir, 2009; Wright, 1978).

*E. coli* abundance does not predict plasmid diversity

We explored whether larger *E. coli* populations could support a greater variety of plasmid replicons. In order to test this, regression was conducted with total *E. coli* abundance against plasmid replicon diversity (Shannon diversity). When the data was pooled by site, *E. coli* abundance was not a good predictor of plasmid replicon diversity ($F_{1, 7} = 2.81, P = 0.14$; Figure 2A). Abundance was also not a good predictor of plasmid diversity within pat, based on an analysis of faecal samples with good sampling of *E. coli* ($N = 7$ or more isolates) ($F_{1,53} = 0.35, P = 0.56$, Figure 2B).

Replicons are not evenly distributed across *E. coli* subgroups

The majority of *E. coli* isolates contained at least one plasmid replicon; just 26% (136 isolates) had no detectable plasmids. Of the 18 possible replicons picked up by the screen 8 were detected. There was one clear dominant type, IncFIB, which was found in 352 isolates (67%). Of the other replicon types identified, IncFIA and IncP were common (> 100 isolates); IncFIC and IncI1 were moderately common (> 30 isolates); and IncB/O,
IncK/B and IncY were rare (< 10 isolates). The full plasmid distribution and H-antigen data sets are available on Dryad (link to be confirmed after acceptance of MS).

Figure 3A highlights the non-random distribution of replicons across *E. coli* subgroups. Flagellin sequences were used to allocate isolates to the four major flagellin gene phylogenetic groups (Ec1a, Ec1b, Ec2, Se2) (Wang *et al.*, 2003). Significantly more replicons than would be expected by chance were found for: IncY in Ec1a; IncP in Ec2; I1, IncFIA and IncFIC in Se2. The replicon distribution for subgroup Se2 is particularly striking, as this is the smallest subgroup (n = 72). Three replicon types are over-represented in Se2, including IncFIC. Significantly reduced carriage of replicons than expected by chance were found for: IncI1, IncFIA and IncFIC replicons in serotype group Ec1a; for IncY and IncP in group Ec1b and for IncP in Se2.

Given that particular groups of *E. coli* serotypes were enriched for different plasmid replicons we might expect that increased within species diversity would lead to increased plasmid diversity. At the site level, H serotype diversity is a reasonably good predictor of plasmid diversity ($F_{1,7} = 6.23$, $R^2 = 0.41$, $P = 0.041$, Figure 3B). This relationship appeared to be independent of the effects of *E. coli* abundance since *E. coli* abundance and H serotype diversity were uncorrelated (Figure 3C, $F_{1,7} = 1.09$, $P = 0.33$). Variation in H serotype diversity, and overall site effects both significantly affected plasmid diversity at the pat level (Figure 3D, $F_{1,51} = 13.8$, $P = 0.0005$, and $F_{8,52} = 4.02$, $P = 0.0009$) and together explained 54% of the variation in plasmid diversity.

Plasmid replicons were also not randomly distributed with respect to each other. We examined whether the three most common replicons, IncP, IncFIA, and IncFIB associated independently. We found that IncP was strongly associated with IncFIB replicons much more commonly than we would expect by chance (Fig 4A, Fisher’s exact test $P < 0.001$), IncP was also found in isolates that tended to lack IncFIA replicons (Fig.
4B, Fisher's exact test $P < 0.001$), while IncF1A was typically found in association with
IncF1B (Fig 4C, Fisher’s exact test $P < 0.001$).

The number of replicons contained within each isolate was also affected by
genetic and ecological factors. H serotype had a significant effect on the number of
replicons found in each isolate (Figure 5A, glm with Poisson errors $df = 28$, $\chi^2 = 156$, $P <$
0.0001). However, this variation was largely due to the influence of three ‘plasmid shy’
serotypes (H4, H27, H28), these three serotypes accounted for more than half of the
plasmid-free isolates (Figure 5). The variation in replicon number could be neatly
described by dividing bacterial serotypes into two categories: ‘plasmid shy’ serotypes
(H4, H27, H28) and ‘plasmid normal’ (all other genotypes) without any significant loss of
explanatory power ($df = 27$, $\chi^2 = 37.5$, $P = 0.086$); in other words there was no significant
variation in replicon counts among the plasmid normal strains. The varying prevalence
of these plasmid shy strains could explain some of the variation in replicon number per
isolate that we saw between sites (glm with Poisson errors $df = 8$, $\chi^2 = 115$, $P < 0.0001$),
in particular note the high prevalence of serotypes H4 and H28 in sites OC and RH
respectively (Figure 1). Nevertheless, site still had an effect on plasmid occupancy when
we considered plasmid normal serotypes only (glm with Poisson errors $df = 8$, $\chi^2 = 58$, $P$
< 0.0001) especially in Site OC which had on average nearly one fewer replicon per
isolate than most other sites (Figure 5B). The number of replicons per isolate fitted a
Poisson distribution fairly well. However, we saw fewer isolates with zero replicons
than we would expect by chance and an excess of isolates with two replicons in
particular (Pearson’s goodness of fit test, $df = 3$, $\chi^2 = 16.8$, $P < 0.001$, Figure 5C). This
pattern was even more marked when we considered the plasmid normal strains only
(Pearson’s goodness of fit test, $df = 3$, $\chi^2 = 40.0$, $P < 0.0001$).
Discussion

Host population structure is recognized as a key determinant of genetic diversity in parasites (Van den Broeck et al., 2014). As plasmids can be considered to act as parasites in some contexts, it follows that the population structure of host cells may also have a large influence of plasmid genetic diversity. Bacterial population structure also has many implications for the maintenance of investment in virulence (Buckling and Brockhurst, 2008; van Leeuwen et al., 2015) and our understanding of how plasmids are maintained in bacterial communities. It is therefore vital to consider the host population structure when examining plasmid populations. In this study, many pats appear to contain E. coli populations with very limited diversity, commonly all isolates from a given pat had the same H-antigen type. Time series studies of cattle and human hosts have shown that hosts tend to be colonized by a single dominant E. coli subtype and a range of less frequent transient subtypes over time (Anderson et al., 2006; Caugant et al., 1981; Jenkins et al., 2003). When high levels of antimicrobial resistance have been found in E. coli in cattle single genotypes can dominate entire farms (Sawant et al., 2007), a pattern that we did not see, potentially because herds reared for conservation management are not subject to strong selective sweeps based on antibiotic exposure. The data in this study fit well with the rarity of mixed infections shown by many bacteria (Balmer and Tanner, 2011), a phenomenon that can be explained by bottlenecks in host colonization (van Leeuwen et al., 2015).

The data on the structure of the host E. coli population allow us to make some inferences about how plasmids persist in this type of community. We found little within species bacterial diversity within hosts, but considerable variation in plasmid carriage within bacterial genotypes within faecal sample (Figure 1, Figure 5). Since E. coli within
For Peer Review Only

one pat (animal) are mostly of one H-antigen type, plasmids must persist via vertical
transmission, or if they are horizontally transferred, the majority of conspecific potential
recipients are likely to be closely related plasmid-free bacteria of the same genotype. A
low diversity *E. coli* population within an animal therefore has implications for the
evolutionary ecology of plasmid persistence; repeated encounters of bacteria and
plasmid genotypes should favour ongoing coevolution between the two. However,
 transient or rare *E. coli* subtypes, or other Gram negative species could introduce new
plasmids to the system via conjugation (Caugant *et al.*, 1981).

One of the most interesting findings in this study is that bacterial host abundance
has little effect on plasmid replicon diversity, whereas bacterial host diversity has a
significant effect. This patterns fitted well with the observation that plasmid replicons
were associated with specific host types (Figure 3), as has been seen in between species
surveys of plasmid carriage (Carattoli, 2009). The strong associations of plasmid
replicons with particular bacterial genotypes indicates that coevolution of plasmids and
hosts through altered conjugation rate, loss of genes or changes in gene expression is
ecologically realistic (Harrison and Brockhurst, 2012; Slater *et al.*, 2008). Experimental
evolution has revealed that the fitness costs associated with the carriage of recently
acquired plasmids can be ameliorated very quickly, either by mutations on plasmids or
the host chromosome (Bouma and Lenski, 1988; Dahlberg and Chao, 2003; Dionisio *et
al.*, 2005; Harrison *et al.*, 2015; San Millan *et al.*, 2014). However, while plasmids can
have low fitness costs in many contexts (Vogwill and Maclean, 2015), these costs can
vary considerably in different host bacteria species (Hall *et al.*, 2015; Riccobono *et al.*,
2015). The variation in fitness cost may be due to different lengths of coevolutionary
history between bacteria and plasmids (Vogwill and Maclean, 2015), but it is not clear
whether this can account for the host associations shown here, within *E. coli*, or across the Enterobacteriaceae (Carattoli, 2009).

Conceivably the opportunities for reducing the costs of carriage are not equal in all genetic contexts. Epistatic interactions between plasmids, that can have significant impacts on fitness costs, may play a role here. Previous work has shown that small plasmids can have substantial effects on the costs of carrying large plasmids, and meta-analysis of sequence data show that large and small plasmids are commonly found together (San Millan *et al.*, 2013). Here, we found evidence for a frequent occurrence of two replicons per bacterial host in ecological communities. While some of these associations may arise because large IncF –type plasmids will often carry both IncF1A and F1B replicons (Villa *et al.*, 2010), the association between IncF1B and IncP replicons was novel. Conversely, antagonistic interactions between IncF1A and IncP, which are known to interfere in other's conjugation potential (Miller *et al.*, 1985) may explain the rare co-occurrence of these two replicon types.

Potentially different bacterial lineages can follow different evolutionary trajectories shaped by initial plasmid complements. In this context, the occurrence of serotypes with consistently low rates of carriage, the plasmid shy strains, is particularly intriguing, especially given the high prevalence of two of them (H4, H28). Investigating the mechanism behind these low rates of carriage would be worthy of future study. The genetic background of *E. coli* plasmid recipients are known to have a considerable, unexplained, effect on conjugation rates (Dionisio *et al.*, 2002). Other possible mechanisms include genetic backgrounds that have higher segregation rates, or which harbour cryptic plasmids which confer partial incompatibility. Site also had a strong influence on rates of plasmid carriage, and this was true even when we excluded the effect of the prevalence of plasmid shy strains. After excluding genotype effects, the site
with the lowest rate of plasmid occupancy (site OC) also had the highest prevalence of plasmid shy bacteria. This is particularly intriguing. In the same way that highly efficient plasmid donors in diverse communities can accelerate plasmid spread (Dionisio et al., 2002), we can speculate that the high prevalence of strains that are ‘plasmid sinks’ might also have an effect on plasmid carriage at the level of the population.

It has been proposed that farm animals may be important carriers of, and reservoirs for, ESBL genes (Carattoli, 2008). However, data on the prevalence of ESBL producing bacteria in food or food-animals is highly variable, ranging from 0.2% to 40% (Liebana et al., 2013). This difference is highlighted by the contrast between data in this study and a previous study of faecal samples from dairy cattle; < 3 % of E. coli were resistant to ampicillin, chloramphenicol or cefotaxime in this study, whereas 48 % and 20 % of E. coli were resistant to ampicillin and chloramphenicol respectively in dairy herds (Sawant et al., 2007). It is of course possible that the E. coli sampled in this study carry resistance to antibiotics other than those tested. However, the lack of ESBL resistance is noteworthy as it may indicate that resistance in livestock is not as prevalent as is supposed. Five of the nine sites in the study were small grazing herds kept on nature reserves. It is highly likely that these cattle have had very low exposure to antibiotics, and the low levels of resistance found suggest that this means of cattle rearing may reduce the prevalence of resistance. The spread of antibiotic resistant organisms through the food chain is considered to be a significant risk to public health, but if resistance in livestock can be managed effectively, this risk may be minimised (Liebana et al., 2013). Another basic issue relating to evolution of antibiotic resistance was the observation of the common variation in plasmid profiles within bacterial genotypes within mammalian hosts. The explanation for this variation and lack of fixation in
plasmid carriage is still not clear. Similarly, we understand little about the striking levels of polymorphism in antibiotic resistance seen in many bacteria (Colijn et al., 2010). Given the common association between plasmid carriage and antibiotic resistance understanding the basic polymorphism in plasmids might be a useful step.

This study presents some fundamental ecology of enteric E. coli and plasmids from cattle. It has found surprisingly low levels of resistance, giving hope that alternative means of cattle rearing may lead to lower prevalence of antibiotic resistance. It has also demonstrated the link between host E. coli diversity and plasmid diversity, and found that E. coli populations are highly structured. Further sequencing of plasmids would allow us to build a fuller picture of the plasmid population, sequencing of bacterial genomes may allow us to identify genetic determinants of plasmid carriage and deep sequencing may also uncover small plasmids that were missed by replicon typing, and enable more accurate analysis of plasmid mobility.

Experimental procedures

Sample collection & isolation of E. coli

Bovine faecal samples (cowpats) were collected from 9 sites across Surrey and Berkshire: four farms or areas of common land (sites A, D, R & W) and five nature reserves with grazing cattle managed by Surrey Wildlife Trust (sites CC, OC, RH, TM and WC). These were chosen because the cattle are extensively grazed on semi-natural grassland and therefore would not have general exposure to antibiotics as part of routine diets. Sampling was conducted in a hierarchical manner: 10 cowpats were sampled from each site and 10 E. coli were screened from each pat (unless numbers were too low, in which case all available pats/ E. coli were screened). Cowpats were
sampled at regular intervals along walking transects in each field, with single pats being
taken randomly from any clustered deposits. *E. coli* survival in cowpats in the field is
robust for up to 30 days, but fresh pats were selected when possible (Van Kessel et al.,
2007). Additional details of sites, site management and mean *E. coli* density is given in
Supplementary table S1.

Approximately 2 g faecal matter was collected from each pat using sterile
collecting tubes with spoon lids (Nunc™, Thermo Scientific, UK). Samples were returned
to the laboratory where they were stored at 10°C until processing. All samples were
processed with 24 hours of collection by homogenisation in 5 ml sterile 0.85% w/v NaCl
followed by centrifugation at 500 g for 2 minutes to pellet the solid material. The
supernatants were diluted 100-fold with sterile 0.85% w/v NaCl, and 100 µl was plated
onto HiCrome™ Coliform Agar containing 5 mg l⁻¹ novobiocin (Sigma, UK). Total coliform
counts were recorded after overnight incubation at 37°C. Ten colonies per cowpat
sample were re-streaked onto HiCrome™ plates to obtain clonal colonies. Overnight
cultures in LB broth were inoculated from these clonal colonies for testing with Kovac’s
Reagent (Sigma). 10 µl Kovac’s Reagent was added to 150 µl overnight culture; a cherry
red colour confirms presence of *E. coli*. Isolated colonies were stored long term in 80%
glycerol at -80°C.

DNA extraction & plasmid replicon typing

Overnight cultures were set up in 2 ml LB broth in 24 well plates. These were
centrifuged at 5500 g for 5 minutes and the pellet resuspended in 0.5 ml molecular
grade H₂O. This suspension was boiled for 10 minutes at 100°C and then centrifuged at
4500 g for 5 minutes. Supernatants were used as templates for replicon typing PCR and
flagellin sequencing; best results were obtained when supernatants were used
immediately. Replicon typing PCR was used to identify large plasmids in isolated *E. coli*. This method comprises 3 multiplex-PCR panels with 18 primer pairs recognising the major plasmid incompatibility groups of the *Enterobacteriaceae* (Carattoli *et al.*, 2005; Johnson *et al.*, 2007). Note that this panel is targeted at large conjugative plasmids and is not therefore expected to cover smaller less well-defined plasmids. PCR reagent concentrations were as follows: 0.4mM dNTPs, 0.625 units Taq (Qiagen) and 2.4 μM of each primer in the multiplex panel. 0.5 μl boiled lysate template was added to each 24.5 μl reaction mix. Cycling conditions were: initial denaturation, 95°C, 5 minutes, followed by 30 cycles of denaturation (95°C, 30 seconds), annealing (60°C, 30 seconds) and extension (72°C, 90 seconds) and a final extension, step of 72°C for 5 minutes.

*E. coli* genotyping

*E. coli* isolates were genotyped by sequencing the variable region of the flagellin gene, *fliC*. This method is an adaptation of previous genotyping methods using whole gene sequencing (Wang *et al.*, 2006) and PCR-RFLP (restriction fragment length polymorphism) of flagellin genes (Machado *et al.*, 2000). The *fliC* gene was amplified using the PCR-RFLP primers *fliC* 1 (5’-CAAGTCATTAATACMAACAGCC) and *fliC* 2 (5’-GACATRTTRGAVACTTCSGT) (Machado *et al.*, 2000). Cycling conditions were as follows: initial denaturation, 95°C, 2 minutes, followed by 30 cycles of denaturation (95°C, 40 seconds), annealing (50°C, 30 seconds) and extension (72°C, 120 seconds) and a final extension step of 72°C for 3 minutes. PCR products were cleaned up using Exo1 (Fisher Scientific) and TSAP (Promega UK, Southampton, UK) both at final concentration 0.045 units μl⁻¹. Clean up mixes were incubated at 37°C for 30 minutes, followed by 80°C for 15 minutes. Sequencing reactions were conducted using BigDye® terminator v3.1 cycle sequencing kit (Life Technologies Ltd. Paisley, UK) using the *fliC* 2 primer only.
Sequencing reactions and clean-ups followed the BigDye recommendations and sequencing runs were conducted by the Department of Zoology, University of Oxford, UK. Trimmed sequences were searched against the whole flagellin gene sequences of Wang et al., (2003) using Geneious (Biomatters 2012). Sequences were assigned a flagellin type (H-antigen type) when they had at least 200 bp of high identity (>95%) with one fliC sequence. Where several matches were made the match with highest identity was selected. This method allowed 97% sequences to be assigned a flagellin type. E. coli H-antigen types were placed into one of four subgroups defined by Wang et al. (2003) (Supplementary table 2). These are based on H-antigen groupings of variable region of the flagellin gene (V). Wang et al. (2003) failed to categorise the H16 group based on its V region sequence, but placed it in the Ec1b group based on the conserved C1 and C2 regions.

Antibiotic susceptibility screening

The susceptibility of E. coli isolates to ampicillin, cefotaxime and chloramphenicol was assessed by disc diffusion assays. MIC breakpoints and interpretation of zone diameters were taken from the British Society for Antimicrobial Chemotherapy Guidelines version 11.1 (May 2012). Susceptibility testing methods followed Andrews (2005), using six antibiotic discs per plate placed at uniform intervals using sterile forceps. Zones of inhibition were measured after overnight incubation at 37°C, using a disc diameter template. Two doses were used for each antibiotic (2 and 10 μg ml⁻¹ ampicillin; 5 and 30 μg ml⁻¹ cefotaxime; 10 and 30 μg ml⁻¹ chloramphenicol) and strains were determined to be resistant if they scored positively with both discs.

Data analysis
All statistical analyses were conducted in R v3.1.2 (R Core Team 2013). Plasmid and H-antigen diversity were calculated using Simpson’s and the Shannon diversity index using the Vegan package in R (Oksanen et al., 2013). Linear regressions were carried out to assess the affect of E. coli abundance on both plasmid and H-antigen diversity, and generalized linear models with Poisson errors used to investigate the number of replicons per bacterial host. Plasmid replicon distribution across subgroups and with respect to each was assessed Pearson’s Chi squared goodness of fit tests and with Fisher’s exact tests. Hierarchical F-statistics were calculated using HierFstat (de Meeus and Goudet, 2007).

Acknowledgments

FM was supported by a joint Royal Holloway and Animal Health and Veterinary Laboratories Agency studentship, BR by a NERC Advanced Research fellowship. Thanks to Tatiana Dimitriu for comments and to Surrey Wildlife Trust for permitting access to several sites.

References


**Supporting Information.** Additional Supporting Information may be found in the online version of the article at the publisher’s website. Table S1 ....
Figure 1. The structure of the *E. coli* population in faecal samples from 9 sites of extensively grazing cattle. A. The frequency distribution of bacterial serotype diversity within samples (= pat). The histogram shows sample counts on the y axis and bins of Simpson's diversity on the x-axis. The inverse Simpson’s diversity index was used to illustrate this distribution as this takes a value of 1 when samples are clonal; most samples were dominated by one serotype. B. A histogram of the variation in number of distinct plasmid profiles within bacterial serotype within each faecal sample, with isolate counts on the y axis and bins of profile richness on the x-axis. In only a minority of samples did isolates of the same serotype have identical plasmid profiles. These data used *N* = 57 pats for which we had at least four isolates of the same serotype. C. The variation in genetic diversity of *E. coli* and distribution of serotype across sites, each serotype has a unique colour, frequency refers to the counts of genotype bacteria at each site. Note that some sites are under-sampled due to low *E. coli* abundance.

Figure 2. Plasmid replicon diversity does not correlate with the abundance of *E. coli* at A. the level of each site study site or B. at the level of the cowpat across all sites. Data in B are log_{10} *E. coli* per pat / sample.

Figure 3. Plasmid replicons are not randomly distributed across *E. coli* serotype groups. A. Mosaic plot of replicon / serotype group association; the sizes of the rectangles correspond to total isolate counts each class; classes with significantly higher than expected counts are coloured in blue, lower than expected counts coloured in red, with the residuals of Pearson’s goodness of fit tests in the legend, *N* = 527 isolates. Note that categorical labels use the abbreviations FA, FB, FC, P etc to indicate IncF1A, IncF1B, IncF1C and Inc P etc replicons. B. Sites with high *E. coli* serotype diversity tend to have
high plasmid replicon diversity. C. *E. coli* serotype diversity is not significantly
correlated with *E. coli* abundance. D. Replicon diversity is positively correlated with *E. coli* serotype diversity within sites. Serotype and replicon diversity data are Shannon
diversity indices for each pat with good sampling, while lines represent the fitted model
with a common slope and different intercepts for each site.

**Figure 4.** Plasmid replicons are not randomly distributed with respect to each other in
*E. coli* isolates. Mosaic plots (see Fig. 3 legend for explanation) show the association
between the three most common replicon incompatibility groups in the 527 isolates. A.
IncP replicons are strongly associated with IncF1B replicons across isolates. B. IncP
replicons are found less often in isolates carrying IncF1A replicons than expected by
chance. C. IncF1A and IncF1B replicons are strongly associated across bacterial
isolates.

**Figure 5.** A. The variation in numbers of plasmid replicons per bacterial isolate
according to *E. coli* H serotype. Frequency refers to counts of isolates with particular
replicon numbers. Three serotypes (indicated by asterisks) have significantly fewer
replicons than the bulk of the genotypes, these are the ‘plasmid shy’ strains; ** and ***
indicate significantly fewer replicons per isolate in *post hoc* comparison at the *P* < 0.01
and *P* < 0.001 respectively. B. Site level variation in numbers of replicons per isolate in
plasmid normal serotypes, *i.e.* excluding plasmid shy strains. Data are means ± 1SE. C.
A histogram of the distribution of the number of replicons per isolate; expected
frequencies were generated from a Poisson distribution with a mean of 1.25 replicons
per host, frequencies are isolate counts (on the y axis), while the x axis represents bins
of number of distinct replicons. We found a significant mismatch between expected and observed frequencies (Pearson’s goodness of fit test: $\chi^2 = 16.8$, $df = 3$, $P < 0.001$).
**E. coli subgroup**

- Pearson residuals:
  - p-value = < 2.22e-16

**Plasmid replications**

- Ec1a
- Ec1b
- Ec2
- Se2

- FB
- P
- FAKB
- none

**H type diversity**

- Plasmid replicon diversity

**Replicon diversity**

- OC
- R
- RH
- A
- CC
- D

**Log₁₀ E. coli density per site**

- TM
- W
- WC

**Bacterial diversity**

- 0.0
- 0.5
- 1.0
- 1.5
- 2.0

**b**

- Plasmid replicon diversity
  - 0.5
  - 1.0
  - 1.5
  - 2.0

- H type diversity
  - 0.5
  - 1.0
  - 1.5
  - 2.0

**c**

- H type diversity
- log₁₀ E. coli density per site

- TM
- W
- WC

- OC
- R
- RH

- A
- CC
- D

**d**

- Replicon diversity
- H type diversity

- 0.0
- 0.5
- 1.0
- 1.5
Pearson residuals:
- p-value = 7.5896e-07
- IncP
- with P
- no P
- with F1B
- no F1B
- -3.6
- -2.0
- 0.0
- 2.0
- 2.6

b) Pearson residuals:
- p-value = 0.00021075
- IncP
- with P
- no P
- with F1A
- no F1A
- -2.9
- -2.0
- 0.0
- 1.6
- 2.0

C) Pearson residuals:
- p-value = 9.4542e-11
- IncF1A
- with F1A
- no F1A
- with F1B
- no F1B
- -4.7
- -4.0
- -2.0
- 0.0
- 2.0
- 3.3