Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of *Marmota himalayana* in Qinghai–Tibet plateau of China

Shan Lu1,2,*, Dong Jin1,2,*, Shusheng Wu3,*, Jing Yang1,2,*, Ruiting Lan4,*, Xiangning Bai1,2, Sha Liu1,2, Qiong Meng1,2, Xuejiao Yuan1,2, Juan Zhou1,2, Ji Pu1,2, Qiang Chen1,2, Hang Dai1,2, Yuanyuan Hu1,2, Yanwen Xiong1,2, Changyun Ye1,2 and Jianguo Xu1,2

*Escherichia coli* is both of a widespread harmless gut commensal and a versatile pathogen of humans. Domestic animals are a well-known reservoir for pathogenic *E. coli*. However, studies of *E. coli* populations from wild animals that have been separated from human activities had been very limited. Here we obtained 580 isolates from intestinal contents of 116 wild Marmot *Marmota himalayana* from Qinghai–Tibet plateau, China, with five isolates per animal. We selected 125 (hereinafter referred to as strains) from the 580 isolates for genome sequencing, based on unique pulse field gel electrophoresis patterns and at least one isolate per animal. Whole genome sequence analysis revealed that all 125 strains carried at least one and the majority (79.2%) carried multiple virulence genes based on the analysis of 22 selected virulence genes. In particular, the majority of the strains carried virulence genes from different pathovars as potential 'hybrid pathogens'. The alleles of eight virulence genes from the Marmot *E. coli* were found to have diverged earlier than all known alleles from human and other animal *E. coli*. Phylogenetic analysis of the 125 Marmot *E. coli* genomes and 355 genomes selected from 1622 human and other *E. coli* strains identified two new phylogroups, G and H, both of which diverged earlier than the other phylogroups. Eight of the 12 well-known pathogenic *E. coli* lineages were found to share a most recent common ancestor with one or more Marmot *E. coli* strains. Our results suggested that the intestinal *E. coli* of the Marmots contained a diverse virulence gene pool and is potentially pathogenic to humans. These findings provided a new understanding of the evolutionary origin of pathogenic *E. coli*.

*Emerging Microbes & Infections* (2016) 5, e122; doi:10.1038/emi.2016.122; published online 7 December 2016

**Keywords:** *Escherichia coli*; evolution; Marmot; pathogenicity; virulence genes

**INTRODUCTION**

*Escherichia coli* is both of a widespread harmless gut commensal and a versatile pathogen of humans, which is estimated to cause more than two million deaths annually through both intestinal and extraintestinal infections.1 On the basis of the mode of pathogenesis and presence of typical virulence factors, there are at least five well recognized pathovars of intestinal pathogenic *E. coli*, including enteropathogenic, enterohaemorrhagic, enteroaggregative and enteroadhesive *E. coli* (EPEC, EHEC, ETEC, EAEC and EIEC, respectively) and two well recognized pathovars of extraintestinal pathogenic *E. coli* (uropathogenic (UPEC) and Neonatal meningitis *E. coli* (NMEC)).1 The major virulence factors for these pathovars are all on mobile elements. EPEC strains carry the LEE pathogenicity island encoding a type III secretion system for effacing and attachment, and the EAF plasmid for adherence. EHEC strains also carry LEE but with two crucial additional elements, Shiga toxin phage(s) and an EHEC plasmid encoding a haemolysin. Shiga toxin *E. coli* (STEC) which may or may not carry LEE is recognized as an extension of EHEC with variable clinical outcomes. ETEC strains contain plasmid borne genes for enterotoxins and specific pili for adhesion. EIEC and *Shigella* strains carry the pINV invasion plasmid and are one pathogenic form as they share the same mode of cell invasion. The typical strains of a pathovar can be defined by one or a few virulence genes.2,3 However, the pathogenicity of UPEC and NMEC is more difficult to define using virulence genes as many genes are implicated.4,5 It seems that UPEC pathogenicity is determined by the total virulence gene content and certain combinations of multiple virulence genes.4,5 However, there are emerging hybrid forms of pathogenic *E. coli* which carry novel combinations of known virulence factors as highlighted by the Shiga toxin producing EAEC O104:H4 which...
caused the German hemolytic uremic syndrome outbreak. There are also UPEC strains carrying intestinal pathogenic virulence factors including EAEC and STEC virulence genes.9

The established view is that pathogenic E. coli was evolved from commensal E. coli in humans or animals through horizontal transfer of virulence genes, resulted into various pathovars to cause clinically and epidemiologically distinctive diseases.10 Domestic animals are natural reservoirs for some human pathogenic E. coli with the typical example of STEC, which normally does not cause disease in these animals. On the other hand, the human ETEC and domestic animal ET}_{E}C strains are differentiated by their adhesion factors and as a result, human ETEC strains normally do not cause disease in domestic animals and vice versa.11 The major human ETEC lineages are estimated to have arisen less than 200 years ago through acquisition of plasmid encoded virulence genes.12

In this study, we found that the majority of the 125 strains of intestinal E. coli from 116 healthy M. himalayana in Qinghai–Tibet plateau, China, carried 22 virulence genes examined including LEE-encoded intimin gene (eaeA), the K88 fimbriae usher gene (faeD), the EAEC heat-stable toxin 1 gene (astA), the EHEC hemolysin gene (ehxA), the Shiga toxin 1 gene (stx1), the InfA homolog adhesin gene (iha), the long polar fimbriae genes (lpfA{O}_{113} and lpfA{O}_{157/O}_{141}), the 60-MDa virulence plasmid adhesin gene (ttxB), P fimbriae usher gene (papC), the brain microvascular endothelial cell invasion gene (ibeA), the cytotoxic distending toxin B gene (cdtB) and the Shigella pathogenicity island encoded gene (shiA).2 Phylogenetic analysis indicated that for eight virulence genes such as eaeA, the alleles found in Marmot E. coli diverged earlier than all known alleles from human and other animal E. coli.

### MATERIALS AND METHODS

**Marmot sampling, virulence gene screening and strain isolation**

The Marmots (M. himalayana) were sampled as part of the animal plague surveillance program conducted in Yushu Tibetan autonomous prefecture, Qinghai province. Of a total of 116 Marmots sampled in June 2012, 38 were from Haxiu county (with an altitude of 4116.5 m above sea level (a.s.l)) and 78 from Anchong site (4525.3 m a.s.l). The human population density in Yushu area in 2013 was 1.43 person/km². The distance from sampling sites, Haxiu (4525.3 m a.s.l). The human population density in Yushu area in June 2012, 38 were from Haxiu county (with an altitude

**Virulence gene sequence phylogenetic analysis**

The virulence gene phylogeny was constructed by aligning the nucleotide sequences of each allele using ClustalW alignment algorithm in MEGA5.2.17 Neighbor-joining trees were constructed in MEGA5.2 using the Kimura 2-parameter substitution model with 1000 bootstrap replications.

**Population structure estimation**

We used population structure analysis software STRUCTURE to identify different E. coli groups and subgroups. We ran 20 000 iterations, following a burn-in period of 10 000 iterations with K (number of groups) between 2 and 11 and chose the most suitable K when the model probability increased dramatically. A cutoff value of q ≥ 0.67 (q: the combined probability of ancestry from each of groups for each of individual isolate) was used to assign individual isolates to their corresponding groups or subgroups.

### RESULTS

**Virulence gene content of intestinal E. coli from M. himalayana**

A total of 116 healthy M. himalayana were captured and sampled from Qinghai–Tibet plateau of China in 2012. Using simple PCR to directly examine the intestinal contents of animals for E. coli virulence genes, we found that all 116 M. himalayana were positive for at least one, and the majority were positive for multiple of the 22 virulence genes, suggesting that all Marmots carried pathogenic E. coli (Table 1).

To test this hypothesis, we obtained five E. coli colonies per animal to get a good representation of the intestinal E. coli population for a total of 580 isolates. Pulse field gel electrophoresis (PFGE) analysis of chromosomal fragments generated by restriction enzyme XbaI digestion showed that the majority of the isolates from the same animal were identical (data not shown). Therefore, one isolate for each PFGE type and at least one isolate for each Marmot were selected for a final set of 125 to represent the 580 isolates. These isolates (hereinafter referred to as strains) were then sequenced using Illumina paired-end 100 bp sequencing and assembled using SOAP de novo (GeneBank JZMB000000000-JZQV00000000).

Genomic analysis revealed that the prevalence of the 22 virulence genes examined in Marmot E. coli strains varied from 0.80% to 76.80% with an average of 24.80%, while the number of virulence genes carried per strain varied from one to nine genes with an average of 5.46 (Figure 1). Unexpectedly, 41.68% of the Marmot E. coli strains contained the porcine ETEC K88 fimbriae gene fadE. Nearly all strains contained one or more of the NMEC and/or UPEC genes. The EIEC/Shigella plasmid encoded virulence gene ipaH was not detected. ShiA which is only present in a proportion of S. flexneri strains was found in a subset of the Marmot E. coli strains. Our results showed that the majority of Marmot E. coli strains (92.8%) contained multiple virulence genes.

Pathogenic E. coli strains can be classified into pathovars by the presence of specific virulence genes. The classification of EPEC, EHEC and STEC is relatively straightforward while for UPEC and NMEC it is
Table 1 Pathovar hybrid patterns of 125 Marmot Escherichia coli strains

| Pathovar hybrid pattern | Number of strains | Percentage of strains |
|-------------------------|-------------------|-----------------------|
| EPEC/ETEC/UEPEC/NMEC/EIEC | 13                | 10.40                 |
| EPEC/ETEC/UEPEC/NMEC    | 10                | 8.00                  |
| EPEC/ETEC/EAEC/UEPEC/NMEC | 9                 | 7.20                  |
| EPEC/EAEC/UEPEC/NMEC    | 9                 | 7.20                  |
| EPEC/EAEC/UEPEC/NMEC/EIEC | 8                | 6.40                  |
| EAEC/UEPEC/NMEC/EIEC    | 8                 | 6.40                  |
| EPEC/UEPEC/NMEC         | 8                 | 6.40                  |
| EAEC/UEPEC/NMEC         | 7                 | 5.60                  |
| EPEC/NMEC               | 6                 | 4.80                  |
| ETEC/UEPEC/NMEC/EIEC    | 5                 | 4.00                  |
| EPEC/ETEC/EAEC/UEPEC/NMEC/EIEC | 5 | 4.00                |
| EPEC/STEC/EIEC          | 4                 | 3.20                  |
| EPEC/EAEC/NMEC          | 3                 | 2.40                  |
| EPEC/NMEC/EIEC          | 3                 | 2.40                  |
| EPEC/ETEC               | 3                 | 2.40                  |
| EAEC/NMEC               | 2                 | 1.60                  |
| ETEC/EAEC/UEPEC/NMEC    | 2                 | 1.60                  |
| ETEC/NMEC               | 2                 | 1.60                  |
| UEPEC/NMEC              | 2                 | 1.60                  |
| EPEC/EHEC               | 1                 | 0.80                  |
| EPEC/EIEC               | 1                 | 0.80                  |
| EPEC/ETEC/EAEC          | 1                 | 0.80                  |
| EPEC/ETEC/EAEC/NMEC     | 1                 | 0.80                  |
| EPEC/STEC/EHEC/EAEC/EIEC | 1             | 0.80                  |
| UEPEC/NMEC/EIEC         | 1                 | 0.80                  |
| ETEC/UEPEC/NMEC         | 1                 | 0.80                  |
| EPEC                 | 7                 | 5.60                  |
| NMEC               | 2                 | 1.60                  |
| Total                          | 125            | 100                   |

Phylogenetic relationships of the Marmot E. coli virulence genes with those of human pathogenic E. coli

We extracted DNA sequences of the 22 virulence genes from Marmot E. coli genomes and inferred their phylogenetic relationship with human E. coli virulence genes. Of 13 genes for which the phylogenetic tree can be rooted with an outgroup, eight were found to contain alleles that were diverged earlier than those from human or other sources, including eaeA, fadD, papA, paa, efa1, tsh, lpfA and pic (Figure 2). Five genes (aslA, ehxA, ibeA, hlyA and cadB) have no earlier diverged alleles observed (Supplementary Figure S1). There was no suitable outgroup for phylogenetic analysis for the remaining nine virulence genes, including astA, espP, hek, etpD, kpsD, iha, toxB, hlyA and shiA (Supplementary Figure S2).

The eaeA gene is located on the LEE pathogenicity island and encodes the intimin adhesin, essential for formation of attaching-and-effacing lesion histopathology of EPEC or EHEC infection.20 Six of the 16 known subtypes21 were found in Marmot E. coli, namely α, β, β2, ε, δ and γ. The majority of Marmot E. coli strains carried either β (22.3%) or β2 (36.4%) allele. When the sequences from Citrobacter rodentium was used as an outgroup, the β-subtype from Marmot E. coli HT2012EP07 was the earliest diverged allele (Figure 2A).

The porcine ETEC K88 fimbriae belong to k-fimbrial clade. The usher gene fadD was used to infer the phylogeny. Using K99 fimbriae usher gene fanD as an outgroup, the fadD allele from Marmot E. coli strain HT2012EP06 diverged the earliest. Interestingly, a human ETEC strain 1766a also carried an fadD,22 which is closely related to the allele from Marmot E. coli strain HT2012DB09 (Figure 2B).

F fimbriae are the principal adherence organelles of UPEC. It is antigenically diverse with 11 known serological variants, due to PapA amino acid sequence variation.23 When using the Proteus Mirabilis mmpA gene encoding the MRJP major fimbrial subunit as an outgroup, a lineage solely composed of Marmot strains diverged the earliest (Figure 2C).

For paa encoding the porcine attaching and effacing associated protein, two divergent lineages are known, with one being associated with EPEC and EHEC, while the other with ETEC.24 The Marmot E. coli HT2012EP04 has two copies of paa: one on chromosome and one on a plasmid, which were grouped into the above two lineages. The chromosomal copy from HT2012ST03 was one of the earlier diverged alleles (Figure 2D).

EHEC factor for adherence 1 (encoded by efa1) is an essential factor for the adherence to intestinal epithelial cells.25 Five Marmot E. coli strains carried efa1. When the sequence from C. rodentium was used as an outgroup, the allele from four Marmot E. coli strains diverged earliest (Figure 2E).

The pic gene encodes a serine protease and has been reported to be present in UPEC, EAEC and Shigella.26 When the sequences from C. rodentium were used as an outgroup, the alleles from Marmot E. coli strains HT2012087 and HT2012ST01 diverged the earliest (Figure 2F).

Long polar fimbriae (encoded by lpfA) is one of the few adhesive factors of EHEC O157:H7 associated with the colonization of the intestine.27 Eight lpfA variants exist, namely lpfA1-1 to lpfA1-5 and lpfA2-1 to lpfA2-3. When Salmonella LT2 was used as an outgroup, lpfA from seven Marmot isolates represented by HT2012015 diverged the earliest (Figure 2G).

The hemagglutinin gene tsh belong to a subclass of IgA protease family.28 When using sequence from Escherichia albertii as an outgroup, the tsh allele from Marmot E. coli strain HT2012007 diverged earlier than that from NMEC and avian pathogenic E. coli (APEC) strains (Figure 2H).
Population structure and evolutionary relationships of Marmot E. coli with human and other animal E. coli

We analyzed the population structure of E. coli with 125 Marmot strains and 355 human and other animal E. coli strains which were selected from 1620 genomes available in the GenBank database, with one genome for each sequence type (ST), and one genome for each host of the same ST if multiple genomes were available. Genome comparison found 540 core genes shared by all 480 genomes. Population structure analysis using STRUCTURE divided the 480 genomes into four groups. Group 1 included phylogroups A and E, each with three subgroups. Group 2 included phylogroups D and F with two subgroups each. Groups 3 and 4 coincided with phylogroup B1 with seven subgroups and B2 with six subgroups respectively. We therefore selected 54 representative strains from these 18 subgroups of four major E. coli groups for phylogenetic analysis. Of the 125 Marmot E. coli strains, 121 fell into the four known phylogroups with 28 in B1, 82 in B2, 10 in E and one in D. None of the Marmot E. coli strains fell into phylogroup A (Figure 3).

Four Marmot E. coli strains HT2012DB09, HT2012095, HT2012039 and HT2012098 were not clustered with any known phylogroups. Three of these outlier strains were assigned into two new groups. HT2012095 and HT2012098 were herein assigned to new phylogroup G, while strain HT2012DB09 was assigned to new phylogroup H. Both phylogroups G and H diverged earlier than the other phylogroups. The fourth outlier strain, HT2012039, is distantly related to phylogroup E and is likely to be a new group (Figure 3). The phylogroup G and H strains are distinctive from Escherichia clade I and clade V.29

We further determined the evolutionary relationships of the Marmot E. coli strains with major human pathogenic E. coli lineages. Phylogenetic analysis of the 125 Marmot E. coli strains and 54 representative E. coli strains from other sources revealed that 12 major lineages of human pathogenic E. coli contained Marmot E. coli strains, namely EHEC1, EHEC2, UPEC1-UPEC3, EPEC1-EPEC3 and NMEC (Figure 4). The lineages EHEC1 and EPEC1 were previously recognized as EHEC O157:H7 and EPEC4 respectively.10 In five lineages, there were Marmot E. coli strains that were diverged earlier than human E. coli strains. When the virulence genes within each lineage were compared, we found four patterns of evolution into human pathovars (Figure 4). However, nearly in all cases, there is also gain of virulence genes that are not typical for a given pathogenic type. Some virulence genes were variably present in a lineage and it is difficult to determine whether they were lost in the human strains or gained by other E. coli strains within the lineage.

The first pattern is an ancestral multi-pathovar hybrid evolving into an e specific human pathovar by gaining new pathovar-specific genes while retaining other pathovar genes. The O157:H7 lineage shared the most recent common ancestor with Marmot E. coli strain HT2012039. The ancestor can be inferred to be an EPEC/NMEC hybrid which acquired EHEC pathogenic specific genes (stx1/stx2, ehxA) to gain a new type of pathogenicity. However it still retained EPEC and NMEC genes. Similarly EPEC1 has an ancestral background of EPEC/UPEC/ NMEC and one of the human strains (CS81-05) gained additional EPEC virulence genes. However the other human strain (C887-10) has gained no additional EPEC-relevant virulence genes, which fits into pattern 4 below.
The second pattern is multi-pathovar hybrids evolving into a specific pathovar with gain of new pathovar-specific genes and at the same time a reduction of other pathovar genes. UPEC1 has evolved from an EPEC/EAEC/NMEC background but lost EAEC genes.

The third pattern is an ancestral single pathovar evolving into the same or different pathovar with gain of pathovar-specific genes. The ancestor of UPEC3 is an NMEC, based on the common virulence properties of the human UPEC strains NA114 and SE15 with Marmot E. coli strains. The gain of UPEC specific genes (for example, papC) by the human strains led to UPEC pathogenicity. The human strains retained NMEC genes and became a pathovar hybrid. The EPEC1 lineage showed a similar pattern. The ancestor carried common EPEC/UPEC/NMEC genes and one of the human EPEC strain gained EPEC specific paa gene during specialization.

The fourth pattern is that the human pathogen shares the same virulence gene profile with Marmot E. coli, suggesting a direct spread from wild animals to humans. The EPEC2/EHEC2 lineage contained four Marmot E. coli strains, all of which belonged to ST342. Three of the four Marmot E. coli strains carried the ehxA gene on a plasmid which is shared by the human AEEC (attaching and effacing E. coli) strain C40-11. Similarly in the EPEC3 lineage, human EPEC strains share the same EPEC/NMEC virulence gene profile as the Marmot E. coli strains. The ancestor of UPEC2 was an EAEC/UPEC/NMEC hybrid and the human UPEC strains retained the same hybrid virulence profile. Similarly the NMEC lineage has an ancestral NMEC background and remained the same when becoming a human pathogen. Some of these lineages (EPEC2/EHEC2, UPEC3) may have been a direct spread from wild animals to humans as they belong to the same STs (for example, the EPEC2/EHEC2 and UPEC3 lineages) and contain the same virulence gene profiles.

DISCUSSION
Domestic and wild animals have been shown to be the reservoir of human pathogenic E. coli. All except EIEC and Shigella have been found in other animals where they normally do not cause disease.
There has been extensive sampling from domestic animals for pathogenic E. coli. However, sampling from wild animals that have been separated from human activities has been very limited. In this study we sampled E. coli from M. himalayana from the Qinghai–Tibet plateau of China which is a relatively pristine environment with little human interference. The Himalayan region is in the high-frigid and meadow-steppe with elevations of 2700–5450 meters and sparse vegetation. The area is sparsely populated with humans, although human settlement traces back to thousands of years ago. The high altitude and harsh climate in the winter make it less suitable for human inhabitation. There are many wild animal species in the region and the species of interest in this study is M. himalayana. Of the 14 Marmota species, only M. himalayana is found in the Himalayan region. M. himalayana is the major animal reservoir of Yersinia pestis in China and the Qinghai–Tibet plateau plague focus is still active with human infections and deaths in recent years. For more than half a century, hunting or trading of M. himalayana has been banned under the national policy for the control of plague, which has helped

Figure 3 The phylogenetic relationship and virulence gene content of intestinal E. coli of M. himalayana and 54 representative E. coli genomes. The tree was constructed with neighbor-joining algorithm-based 540 core genes. The presence of pathovar-specific genes in each strain was indicated by colored boxes with the color scheme shown. A strain is positive for a specific pathovar if at least one pathovar-specific gene is present. White space represents absence of pathovar-specific genes for a given pathovar in that strain. Strains from Marmots were highlighted in red. Phylogroups were labeled in the inner circle as well as by coloring of the branches.
maintain the pristine environment. The *E. coli* we sampled was more likely to be the native population colonizing *M. himalayana*, rather than ingression from humans or domesticated animal species. We showed that all intestinal *E. coli* strains of *M. himalayana* carried virulence genes and are potentially pathogenic to humans. Importantly, the Marmot *E. coli* carried an ancient virulence gene pool as alleles of eight key virulence genes from these *E. coli* strains diverged earlier than alleles of *E. coli* from other sources.

The Marmot *E. coli* strains were not a recent ingression from humans or domestic animal hosts based on the antibiotic resistance profiles. There is a higher prevalence of antibiotic resistance in human *E. coli* population than that of wild animals. Our Marmot *E. coli* strains were sensitive to all 23 antibiotics tested (data not shown), including penicillins (ampicillin and piperacillin), β-lactam/β-lactamase inhibitor combinations (amoxicillin-clavulanic acid and ampicillin-sulbactam), cephalosporins (cefepime, cefotaxime, ceftriaxone, cephalothin and cefuroxime), monobactams (aztreonam), carbapenems (imipenem and meropenem), aminoglycosides (gentamicin, kanamycin and streptomycin), tetracycline, fluoroquinolones (ciprofloxacin, norfloxacin and levofloxacin), nalidixic acid, trimethoprim-

---

**Figure 4**  The evolution of major human pathogenic lineages of *E. coli* relative to their closely related Marmot *E. coli* strains. Key non-Marmot *E. coli* strains are given by name and Marmot *E. coli* strains were highlighted with a red dot. Outside the circle in the different shaded circles are a detailed depiction of the pathogenic *E. coli* lineages with gain and loss of genes. Gain of gains is marked by a black arrow, while loss of genes is marked by a vertical green bar. *Escherichia* Clades I–V strains were used as an outgroup.29
sulfamethoxazole, chloramphenicol and nitrofurantoin. Therefore, the Marmot E. coli we sampled was most likely the natural native population of the intestinal flora of the Marmot. The chance of contamination by human E. coli was also low as human population density is low at 1.43 persons/km². The distance from sampling sites to nearest human residence village is 10 km to 15 km.

In sharp contrast to what have been observed in human pathogenic E. coli that most strains carried pathovar-specific virulence genes, the Marmot E. coli strains have been found to be mostly hybrid forms of carrying virulence genes from different pathogens. It has been considered to be the exception rather than the rule for hybrid forms of E. coli with the emergence of STEC/EAEc O104:H4 that caused the German outbreak. Other EHEC/EAEc hybrid strains, STEC/ETEC, UPEC/STEC, and UPEC/EAEc forms have also been reported recently. These reports suggest that such hybrid forms exist as human pathogens but are relatively infrequent. In comparison, most of the Marmot E. coli strains are hybrids of two or more forms. Although assigning a strain to a pathovar based on one or a few genes is very simplistic and the hybrid forms may not cause clinical disease in humans in multiple forms, our results showed a complex picture of virulence in E. coli and high fluidity of virulence genes in the Marmot E. coli population. It will be interesting to examine E. coli from other wild animals to see whether this is a general picture. Our data also indicated that during their evolution to become human pathogens, modern pathogenic E. coli may have lost virulence genes not needed for the specific type of pathogenicity. However it should also be noted that humans may be an accidental host for the various forms of pathogenic E. coli. These genes may have a very different role for the E. coli living in its other hosts such as Marmots or the environment.

Our data suggest that some modern human pathogenic E. coli may have been derived from ancient pathogens by specialization rather than acquisition of a full set of virulence genes by commensal non-pathogenic E. coli to become single mode pathogens. Many pathogens such as O157:H7 have an ancestral pathogenic background with their essential virulence genes already present in their ancient progenitor. Our findings are supported by earlier studies that suggest virulence gene acquisition and expression by human pathogenic E. coli require certain phylogenetic background and that extraintestinal virulence is an ancestral character within the B2 phylogenetic group. Our results crystallize the previous findings and revealed different patterns of derivation of modern human pathogens from the primordial pathogens.

The distribution of E. coli phylogroups has been extensively studied with differential prevalence of the phylogroups in different hosts or human population of different regions. The Marmot E. coli strains predominantly belonged to B2 group with 67.8% and then B1 group with 23.1%. Studies of Australian native animal also showed wild animals harbor predominantly B2 strains with over 30%, while domestic animals harbor predominately B1 strains. E. coli phylogroups D, E, B1 and B2 diverged about 10 million years ago and underwent major expansion about five million years ago. This timeline is within the time frame when Marmots diverged from squirrels about six million years ago, suggesting that E. coli had evolved into these major phylogroups before they colonized Marmots. In addition, we identified two new phylogroups G and H with the latter being the earliest diverged phylogroup.

This study found that none of Marmot E. coli strains fell into phylogroup A. In contrast, phylogroup A is present in high frequency in E. coli from human fecal samples, varying from 20% to 77% depending on the geographical region. Phylogroup A is known to have diverged the latest among the phylogroups. By virulence gene content, phylogroup A contained the least number of virulence factors. Indeed the well-known commensal E. coli strains in phylogroup A, HS, K-12 and BL21, contained none, one (aslA) and three (aniA, kpsD and shiA) of the 22 virulence genes, respectively. It seems that phylogroup A has undergone virulence gene shedding. Therefore, it is likely that human commensal E. coli may be an ultimately evolved rather than the primitive form of E. coli.

In conclusion, Marmot E. coli carried many virulence genes in a mixed virulence gene pool and hybrid pathogenic forms were found which have the potential to cause disease in humans. Two new phylogroups, H and G, both of which diverged earlier than the other phylogroups were identified. The Marmot E. coli population seemed to carry an ancient virulence gene pool. Eight of the 12 well-known pathogenic E. coli lineages shared a most recent common ancestor with one or more Marmot E. coli strains. These findings provided a new understanding of the evolutionary origin of pathogenic E. coli.

ACKNOWLEDGEMENTS
This work was supported by grants from the National Natural Science Foundation of China (81290340, 81290345 and 812111251) and from the Ministry of Science and Technology, People’s Republic of China (Mega Project of Research on The Prevention and Control of HIV/AIDS, Viral Hepatitis Infectious Diseases, 2013ZX10004-101; the National Key Research and Development Plan, 2016YFC1201903).

1 Nataro JP, Kaper JB. Diarrheagenic Escherichia coli. Clin Microbiol Rev 1998; 11: 142–201.
2 Muller D, Greune L, Heusipp G et al. Identification of unconventional intestinal pathogenic Escherichia coli isolates expressing intermediate virulence factor profiles by using a novel single-step multiplex PCR. Appl Environ Microbiol 2007; 73: 3380–3390.
3 Vidal M, Kruger E, Duran C et al. Single multiplex PCR assay to identify simultaneously the six categories of diarrheagenic Escherichia coli associated with enteric infections. J Clin Microbiol 2005; 43: 5362–5365.
4 Kohler CD, Dobrindt U. What defines extraintestinal pathogenic Escherichia coli? Int J Med Microbiol 2011; 311: 642–647.
5 Johnson JR, Russo TA. Molecular epidemiology of extraintestinal pathogenic (uropathogenic) Escherichia coli. Int J Med Microbiol 2005; 295: 383–404.
6 Rohde H, Qin J, Cui Y et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. N Engl J Med 2011; 365: 718–724.
7 Prager R, Lang C, Aurass P et al. Two novel EHEC/EAEc hybrid strains isolated from human infections. PLoS One 2014; 9: e95379.
8 Nyholm O, Halkialahti J, Wiklund G et al. Comparative Genomics and Characterization of Hybrid Shiga-toxinogenic and Enterotoxigenic Escherichia coli (STEC/ETEC) Strains. PLoS One 2015; 10: e0135936.
9 Toval F, Kohler CD, Vogel U et al. Characterization of Escherichia coli isolates from hospital inpatients or outpatients with urinary tract infection. J Clin Microbiol 2014; 52: 407–418.
10 Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. Parallel evolution of virulence in pathogenic Escherichia coli. Nature 2000; 406: 64–67.
11 Fleckner JM, Hardwidge PR, Munson GP et al. Molecular mechanisms of enterotoxigenic Escherichia coli infection. Microbes Infect 2010; 12: 89–98.
12 von Mentzer A, Connor TR, Wieler LH et al. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat Genet 2014; 46: 1321–1326.
13 Bai X, Zhao A, Lan R et al. Shiga toxin-producing Escherichia coli in yaks (Bos grunniens) from the Gringhai-Tibetan Plateau, China. PLoS One 2013; 8: e59537.
14 Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. J BMC Bioinformatics 2004; 5: 113.
15 Tiedje JD. Phylogenetic analysis using PHYLIP. Methods Mol Biol 2000; 132: 243–258.
16 Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 2007; 23: 127–128.
17 Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 2008; 9: 299–306.
18 Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003; 164: 1567–1587.
19 Kaper JB, Nataro JP, Mobley HL. Pathogenic Escherichia coli. Nat Rev Microbiol 2004; 2: 123–140.
20 Jerse AE, Kaper JB. The eae gene of enteropathogenic Escherichia coli encodes a 94-kilodalton membrane protein, the expression of which is influenced by the EAF plasmid. Infect Immun 1991; 59: 4302–4309.

21 McGraw EA, Li J, Selander RK, Whittam TS. Molecular evolution and mosaic structure of alpha, beta, and gamma intimins of pathogenic Escherichia coli. Mol Biol Evol 1999; 16: 12–22.

22 Del Canto F, Botkin DJ, Valenzuela P et al. Identification of Coli Surface Antigen 23, a novel adhesion of enterotoxigenic Escherichia coli. Infect Immun 2012; 80: 2791–2801.

23 Johnson JR, Stell AL, Scheutz F et al. Analysis of the F antigen-specific papI alleles of extraintestinal pathogenic Escherichia coli using a novel multiplex PCR-based assay. Infect Immun 2000; 68: 1587–1599.

24 Bruant G, Zhang Y, Garneau P et al. Two distinct groups of porcine enteropathogenic Escherichia coli strains of serogroup O45 are revealed by comparative genomic hybridization and virulence gene microarray. BMC Genomics 2009; 10: 402.

25 Aset JE, Bruant G, Brousseau R et al. Identification of virulence genes linked with diarrhea due to atypical enteropathogenic Escherichia coli by DNA microarray analysis and PCR. J Clin Microbial 2006; 44: 3703–3711.

26 Ruiz-Perez F, Nataro JP. Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity, and role in virulence. Cell Mol Life Sci 2014; 71: 745–770.

27 Torres AG, Blanco M, Valenzuela P et al. Genes related to long polar fimbriae of pathogenic Escherichia coli strains as reliable markers to identify virulent isolates. J Clin Microbiol 2009; 47: 2442–2451.

28 Frommel U, Lehmann W, Rodiger S et al. Adhesion of human and animal Escherichia coli strains in association with their virulence-associated genes and phylogenetic origins. Appl Environ Microbiol 2013; 79: 5814–5829.

29 Luo C, Walk ST, Gordon DM et al. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci USA 2011; 108: 7200–7205.

30 Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nat Rev Microbiol 2010; 8: 207–217.

31 Pupo GM, Lan R, Reeves PR, Bavister P. Population genetics of Escherichia coli in a natural population of native Australian rats. Environ Microbiol 2000; 2: 594–610.

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

33 Lesca M, Clermont O, Woerther PL et al. Commensal Escherichia coli strains in Guiana reveal a high genetic diversity with host-dependent population structure. Environ Microbiol Rep 2013; 5: 49–57.

34 Cui Y, Yu C, Yan Y et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci USA 2013; 110: 577–582. Chinese.

35 Shen EL. [Human plague during 1979-1988 in China and strategy of its control]. Zhonghua Liu Xing Bing Xue Za Zhi 1990; 11: 156–159.

36 Kaper JB, Nataro JP, Mobley HL. Pathogenic Escherichia coli. Nat Rev Microbiol 2004; 2: 17.

37 Le Gall T, Clermont O, Gouriou S et al. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. Mol Biol Evol 2007; 24: 2373–2384.

38 Levin BR. The evolution and maintenance of virulence in microparasites. Emerg Infect Dis 1996; 2: 93–102.

39 Escobar-Paramo P, Clermont O, Blanc-Potard AB et al. A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli. Mol Biol Evol 2004; 21: 1085–1094.

40 Wirth T, Falush D, Lan R et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol 2006; 60: 1136–1151.

41 Li T, Wang J, Su W, Nie W, Yang F. Karyotypic evolution of the family Sciuridae: inferences from the genome organizations of ground squirrels. Cytogenet Genome Res 2006; 112: 270–276.

42 Thomas WK, Martin SL. A recent origin of marmots. Mol Phylogenet Evol 1993; 2: 330–336.

43 Bolhín J, Brynilsrud OB, Sekse C, Snipen L. An evolutionary analysis of genome expansion and pathogenicity in Escherichia coli. BMC Genomics 2014; 15: 882.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2016

Supplementary Information for this article can be found on the Emerging Microbes & Infections website (http://www.nature.com/emi)