No evidence for a bovine mastitis *Escherichia coli* pathotype

Andreas Leimbach1,2,3#, Anja Poehlein2, John Vollmers4, Rolf Daniel2, Ulrich Dobrindt1,3#

1: Institute of Hygiene, University of Münster, Münster, Germany
2: Department of Genomic and Applied Microbiology, Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University of Göttingen, Göttingen, Germany
3: Institute for Molecular Infection Biology, Julius-Maximilians-University of Würzburg, Würzburg, Germany
4: Leibniz Institute DSMZ, German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

Andreas Leimbach: aleimba@gmx.de, https://orcid.org/0000-0001-9077-1041
Anja Poehlein: apoehle3@gwdg.de, https://orcid.org/0000-0002-2473-6202
John Vollmers: jov14@dsmz.de, https://orcid.org/0000-0001-5980-5178
Rolf Daniel: rdaniel@gwdg.de, https://orcid.org/0000-0002-8646-7925
Ulrich Dobrindt: dobrindt@uni-muenster.de, https://orcid.org/0000-0001-9949-1898

Running title: Comparative genomics of *E. coli* bovine mastitis and fecal commensal isolates

corresponding authors
Address correspondence to A.L. (email: aleimba@gmx.de) or U.D. (email: dobrindt@uni-muenster.de).

Abstract

Background: *Escherichia coli* bovine mastitis is a disease of significant economic importance in the dairy industry. Molecular characterization of mastitis-associated *E. coli* (MAEC) did not result in the identification of common traits. Nevertheless, a mammary pathogenic *E. coli* (MPEC) pathotype has been proposed suggesting virulence traits that differentiate MAEC from commensal *E. coli*. The present study was designed to investigate the MPEC pathotype hypothesis by comparing the genomes of MAEC and commensal bovine *E. coli*.
Results: We sequenced the genomes of eight *E. coli* isolated from bovine mastitis cases and six fecal commensal isolates from udder-healthy cows. We analyzed the phylogenetic history of bovine *E. coli* genomes by supplementing this strain panel with eleven bovine-associated *E. coli* from public databases. The majority of the isolates originate from phylogroups A and B1, but neither MAEC nor commensal strains could be unambiguously distinguished by phylogenetic lineage. The gene content of both MAEC and commensal strains is highly diverse and dominated by their phylogenetic background. Although individual strains carry some typical *E. coli* virulence-associated genes, no traits important for pathogenicity could be specifically attributed to MAEC. Instead, both commensal strains and MAEC have very few gene families

enriched in either pathotype. Only the aerobactin siderophore gene cluster was enriched in commensal *E. coli* within our strain panel.

Conclusions: This is the first characterization of a phylogenetically diverse strain panel including several MAEC and commensal isolates. With our comparative genomics approach we could not confirm previous studies that argue for a positive selection of specific traits enabling MAEC to elicit bovine mastitis. Instead, MAEC are facultative and opportunistic pathogens recruited from the highly diverse bovine gastrointestinal microbiota. Virulence-associated genes implicated in mastitis are a by-product of commensalism with the primary function to enhance fitness in the bovine gastrointestinal tract. Therefore, we put the definition of the MPEC pathotype into question and suggest to designate corresponding isolates as MAEC.

Keywords: *E. coli*, pathotype, bovine mastitis, commensals, comparative genomics, phylogeny, virulence, genomic diversity

Background

Bovine mastitis is a common disease in dairy cows with a global economic impact [1]. Mastitis is an inflammation of the cow udder mostly triggered by the invasion of pathogenic bacteria, leading to reduced milk production and quality. *Escherichia coli* is a major causative agent involved in acute bovine mastitis with a usually fast recovery rate. However, in extreme cases *E. coli* mastitis can lead to severe systemic clinical symptoms like sepsis concurrent with fever [2,3]. Occasionally, an infection with *E. coli* results in a subclinical and persistent pathology [4,5]. Traditionally, *E. coli* associated with intramammary infections are considered to be environmental opportunistic pathogens [6]. Thus, the outcome and severity of *E. coli* mastitis was mainly attributed to environmental factors and the cow's innate immune response reacting to pathogen-associated molecular patterns (PAMPs) (most prominently lipopolysaccharide, LPS) rather than the virulence potential of the invading strain [7]. Intramammary infusion of purified LPS induces udder inflammation symptoms similar, yet not identical, to *E. coli* invasion [7,8]. The bovine gastrointestinal tract is a natural reservoir for commensal and pathogenic *E. coli* of high phylogenetic and genotypic diversity with the putative ability to cause mastitis [9]. Nevertheless, it was proposed that various genotypes of *E. coli* with specific phenotypes are better suited to elicit mastitis than others [3,10,11].

E. coli is a highly diverse species with commensal as well as pathogenic strains, which can colonize and persist in humans, animals, as well as abiotic environments [12–14]. The population history of *E. coli* is largely clonal and can be structured into six major phylogenetic groups: A, B1, B2, D1, D2, and E [14–16], some publications also designate phylogroups D1 and D2 as D and F, respectively. These phylogroups have a different prevalence in various human and animal populations, but no host-restricted strains could be identified [14]. Pathogenic *E. coli* isolates are classified in different pathotypes according to the site of infection, clinical manifestation of the disease, and virulence factor (VF) repertoire. The group of intestinal pathogenic *E. coli* (IPEC) includes well-known pathotypes like enterohaemorrhagic *E. coli* (EHEC), enteroaggregative *E. coli* (EAEC), enteropathogenic *E. coli* (EPEC), and enterotoxigenic *E. coli* (ETEC). The most prominent extraintestinal pathogenic *E. coli* (ExPEC) pathotypes are uropathogenic *E. coli* (UPEC), newborn meningitis-associated *E. coli* (MNEC),

and avian pathogenic *E. coli* (APEC) [17–22]. In contrast to IPEC, which are traditionally considered to have a conserved VF repertoire, ExPEC are derived from different phylogenetic lineages and have variable VF content. Various combinations of VFs can lead to the same extraintestinal disease outcome, which solely defines an ExPEC pathotype [15,19,21,23]. However, many of these virulence-associated genes are also present in commensal strains and can rather be considered encoding fitness factors (FFs), that enable or facilitate initial colonization and the establishment of an infection. These FFs have primarily evolved for gastrointestinal colonization as well as persistence, and the ability to cause extraintestinal disease is a coincidental by-product of commensalism. As a consequence, ExPEC are considered to be facultative pathogens that are recruited from the normal intestinal microbiota [14,21,24,25].

The broad spectrum of *E. coli* lifestyles and phenotypes is a result of the underlying genomic plasticity of *E. coli* strains [21]. Only up to 60% of each genome is shared by all isolates, the so-called core genome [26]. This core genome codes mainly for essential housekeeping functions. The remaining genomic information represents the flexible genome with highly variable presence or absence in individual strains. It includes genes for specific habitat adaptations or environmental conditions, and is the basis for the phenotypic diversity of *E. coli* [15,21,27]. The flexible genome consists largely of mobile genetic elements (MGEs), like plasmids, genomic islands (GI), and phages, which facilitate horizontal gene transfer (HGT) and are the driving forces for microbial diversity, evolution, and adaptation potential [28].

Despite the proposal of a mammary pathogenic *E. coli* (MPEC) pathotype [3] and extensive research, no common genetic traits or VFs have been identified for *E. coli* mastitis isolates, so far [11,29–31]. Recently, several publications analyzed *E. coli* genomes from intramammary infections, thereby expanding the method spectrum by comparative genomics approaches [32–35]. All of these studies identified various MPEC regions and genes with different specificity criteria and significance, many of which are not considered to be classical VFs (or even encode for unknown hypothetical functions), but also genes coding for a type VI secretion system (T6SS), LPS biosynthesis, biofilm association, metabolic functions, and the ferric iron(III)-dicitrate (Fec) uptake system. However, the studies could mostly not agree upon a common set of putative VFs, except for the Fec siderophore system. Also, these studies suffer from small genome sample size constraints, lack of phylogenetic diversity, and/or did not include commensal bovine *E. coli* comparator strains of suitable phylogenetic and genotypic diversity. So far, depending on the study, no or only one bovine commensal *E. coli* isolate has been included in these corresponding analyses [32–35].

We wanted to advance upon the previous studies by analyzing a strain panel of phylogenetic and genomic diversity comparable to *E. coli* from the bovine habitat, especially by including fecal commensal isolates from udder-healthy cows. This enables our main goal, to characterize genetic traits which define mastitis-associated *E. coli* (MAEC) in comparison to non-pathogenic commensals, while keeping track of their phylogenetic background. Putative VFs important for bovine mastitis pathogenesis should be present in the majority of mastitis isolates, regardless of phylogroup, and absent in commensals. We collected a large *E. coli* VF panel from different

pathotypes for detailed candidate gene and gene composition analyses. By finishing two MAEC genomes we made it possible to analyze MGEs and evaluate their role in HGT as well as virulence of MAEC and commensal isolates. Finally, several studies suggested that mastitis virulence might have evolved in separate *E. coli* lineages and phylogroups in parallel [10,11,34,35], which might involve different virulence traits and strategies. Thus, we investigated the distribution of three putative phylogroup A MPEC-specific regions from Goldstone *et al.* [35] within the phylogroups of our strain panel for pathotype association.

Results and discussion

General genome characteristics

We compiled a strain panel of eight MAEC and six fecal commensal strains and supplemented it with the genomes from eleven reference strains from public databases (Table 1). The reference strains are composed of eight MAEC, two fecal commensal strains, and one milk commensal strain. This is the first study which investigates *E. coli* genomes in relation to bovine mastitis including two finished genomes, MAEC 1303 and ECC-1470. The genomes of two phylogroup A MAEC (VL2732 and VL2874) and a phylogroup B1 fecal commensal strain (K71) from published works were not available at the onset of this study [32,34]. For the same reason, the genomes of 66 *E. coli* Reference collection (ECOR) phylogroup A MAEC from a recent study were not included [35].

The assembly statistics of the draft genomes indicate a suitable quality for our analysis purposes with 24 to 290 contigs and N50 values ranging from 79 to 358 kb for contigs larger than 500 bp (Additional file 1: Table S1). There are four exceptions: First, the genome of commensal reference strain AA86 has gone through multiple gap closure steps and has only five contigs with an N50 of 2,860 kb [36]. Two of these five contigs are plasmids, making AA86 the only strain with resolved plasmid sequences in the strain panel in conjunction with the finished 1303 and ECC-1470 genomes [37]. Second, the three MAEC reference draft genomes D6-117.29, ECA-727, and ECA-O157 are highly fragmented with more than 500 contigs each. However, their coding percentage and overall CDS (coding DNA sequence) numbers are in the range of other *E. coli* genomes and thus were included in the strain panel (Additional file 2: Table S2).

Serotypes were predicted *in silico* (Table 1), but could not be determined unambiguously for several draft genomes. Nevertheless, none of the analyzed strains displayed identical serotypes (except non-typable MAEC 131/07 and 3234/A). Thus, a correlation between certain serotypes and MAEC was not detected. This has already been observed earlier for *E. coli* mastitis isolates [10]. Furthermore, divergent serotypes suggest a phylogenetic diversity of the bovine-associated *E. coli* in the strain panel.

Bovine-associated *E. coli* are phylogenetically highly diverse and dominated by phylogroups A and B1

The detected serotypes already indicated a high phylogenetic diversity in the strain panel. In order to obtain a more detailed view of the phylogenetic relationship of the strains, we calculated a core genome phylogeny based on a multiple whole genome nucleotide alignment (WGA) with 39 reference *E. coli* strains, four *Shigella* spp., and one *Escherichia fergusonii* strain as an outgroup. The filtered core genome WGA had a final alignment length of 2,272,130 bp, which is approximately 46% of the average *E. coli* genome size in the strain panel (4,987,267 bp). The resulting *E. coli* population structure resolved the phylogenetic lineages described for *E. coli*, A, B1, E, D1, D2, and B2, with high bootstrap support values and is in consensus with earlier studies [14–16,38].

Table 1 Characteristics of the bovine-associated *E. coli* strain panel. Strains sequenced in this study are highlighted in bold.

Strain	Pathotype	Phylogroup (ST, CC)	Serotype	CDS	Contigs	Reference
1303	MAEC	A (10, 10)	O70:H32	4734	finished	This study, [37]
131/07	MAEC	A (10, 10)	Ont:H39	5123	270	This study, [39]
2772a	MAEC	B1 (156, 156)	O174:H28	4621	93	This study, [39]
3234/A	MAEC	A (10, 10)	Ont:H39	5211	290	This study, [39]
AA86	fecal commensal	B2 (91, 1876)	O39:H4	4627	5	[36]
D6-113.11	MAEC	E (4175, 4175)	O80:H45	4750	89	[34,40]
D6-117.07	MAEC	A (10, 10)	O45:H11	4477	51	[34,40]
D6-117.29	MAEC	A (10, 10)	O28ac/O42:H37	4732	980	Direct submission
ECA-727	MAEC	A (10, 10)	O99:H9	4779	539	[33]
ECA-O157	MAEC	A (398, 398)	O29:H27	4434	1173	[33]
ECC-1470	MAEC	B1 (847, 847)	Ont:H2	4506	finished	This study, [37]

ECC-Z	MAEC	A (10, 10)	O74:H39	4600	24	[33]
MPEC4839	MAEC	A (10, 10)	O105:H32	4502	124	This study, [39]
MPEC4969	MAEC	B1 (1125, 161)	O139:H19	4468	130	This study, [39]
O157:H43 T22	milk commensal	B1 (155, 58)	O157:H43	4792	64	[41–43]
O32:H37 P4	MAEC	A (10, 10)	O32:H37	4581	72	[32,44]
P4-NR	MAEC	B1 (602, 446)	O15:H21/H54	4569	107	Direct submission
RiKo 2299/09	fecal commensal	B1 (448, 448)	O8/O160:H8	4587	129	This study, [39]
RiKo 2305/09	fecal commensal	B1 (410, 88)	O8:H21	4429	123	This study, [39]
RiKo 2308/09	fecal commensal	A (167, 10)	O9a/O89:H9	4685	186	This study, [39]
RiKo 2331/09	fecal commensal	B1 (1614, NA)	Ont:H23	4350	59	This study, [39]
RiKo 2340/09	fecal commensal	A (167, 10)	O89:H9	4568	204	This study, [39]
RiKo 2351/09	fecal commensal	B1 (88, 88)	O21:H4	4931	252	This study, [39]
UVM2	MAEC	A (1091, 10)	O53:H10	4614	149	This study, [39]
W26	fecal commensal	B1 (1081, 533)	O45:H14	4865	165	[45]

The 25 *E. coli* genomes of the bovine-associated strain panel were mostly associated with phylogroups A and B1 (13 and 10, respectively; Table 1). Most of the MAEC (11/16, 69%) belong to phylogroup A and the majority of commensal strains to group B1 (6/9, 67%). MAEC

D6-113.11 and commensal AA86 are the exception to the rule by being associated with phylogroups E and B2, respectively. All phylogenetic group affiliations of the included reference strains were in accordance to their source publications: commensal AA86 (B2); MAEC D6-113.11 (E) and D6-117.07 (A); MAEC ECA-727, ECA-O157, and ECC-Z (all A); MAEC O32:H37 P4 (A); commensal O157:H43 T22 (B1); commensal W26 (B1). Phylogroup A is traditionally associated with commensal strains, its sister taxon B1 is associated with commensals and different IPEC including ETEC, EAEC, and EHEC [20–22]. These two phylogroups represent the youngest within the *E. coli* phylogeny and branch most distally from the root. ECOR phylogroup E includes the genetically closely related O157:H7 EHEC and O55:H7 EPEC, which both originate from a common ancestor [46]. Interestingly, strain O157:H43 T22, even though belonging to the O157 serotype, is not a member of phylogroup E, but of group B1 [43], providing an example for horizontal transfer of O-antigen genes. *Shigella dysenteriae* Sd197 is sometimes classified as ECOR phylogroup E, as it is a member of the same monophyletic clade [47]. Finally, phylogroup B2 is the most diverse phylogroup, based on nucleotide and gene content. This group also includes most of the ExPEC, like UPEC, APEC, and MNEC [14,18,20,22,48]. However, with the accumulation of *E. coli* sequencing data, the traditional association of phylogroups with pathotypes have softened, as many pathotypes were shown to have emerged in parallel in different lineages [16,21,22].

The phylogenetic placement of the strains is in agreement with previous studies where MAEC and bovine commensals were also enriched in phylogroups A and B1, while other phylogroups play only a minor role [9,11]. Depending on the study and the respective analyzed strain panel, MAEC isolates are either more common in phylogroup A [6,30,31] or phylogroup B1 [11,34,49,50]. Blum and Leitner also showed that MAEC are more closely associated with phylogroup A than environmental (bovine fecal) *E. coli* isolates, but the majority of both isolate groups was from the B1 lineage [11]. Also, the WGA phylogeny shows that bovine MAEC and commensals do not cluster together, but rather originate from diverse lineages within phylogroups (Figure 1) [11,34,35]. The discrepancies of MAEC phylogroup associations between the previous studies might be a result of country-specific differences or differences in sampling and phylotyping techniques. The Clermont triplex PCR for phylogroup assignment, which was extensively used in almost all of these studies, has several major drawbacks such as considering only phylogroups A, B1, B2, and D1. Also, it is difficult to assign strains to phylogroup A with this method, because phylogroup A classification is determined by the absence of a PCR product [34,51,52]. An improvement of the method, a quadruplex PCR, has been proposed in 2013 [53] but, like other methods with more discriminatory power such as multi-locus sequence typing (MLST), it is not widely used yet. This is probably the main reason for the past oversight of MAEC in phylogroup E, which was only recently remedied by Kempf *et al.* 2016 with an MLST approach [34].

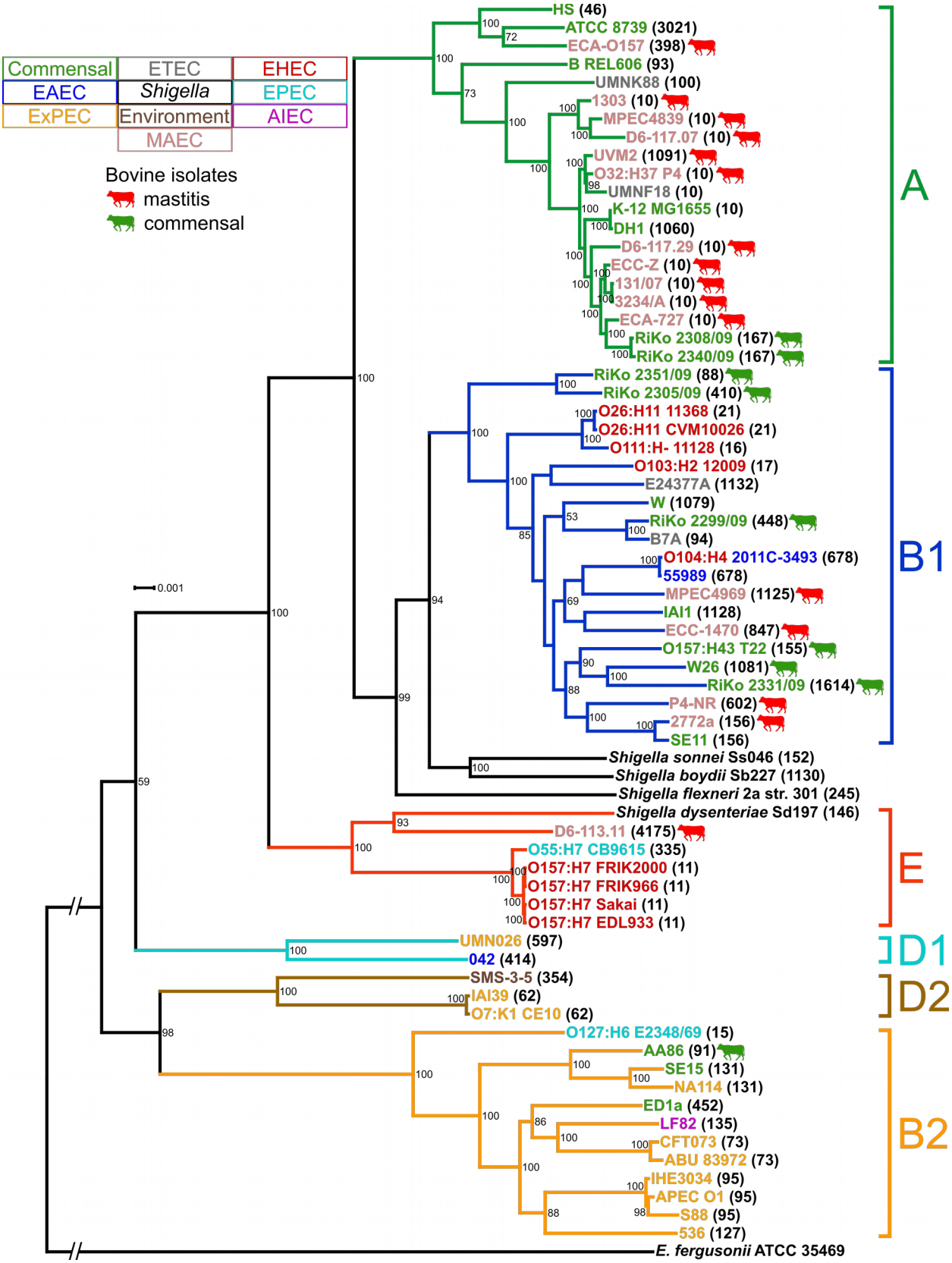

Figure 1 Whole genome alignment phylogeny of bovine-associated and reference *E. coli* strains. The phylogeny is based on a whole core genome alignment of 2,272,130 bp. The best scoring maximum likelihood (ML) tree was inferred with RAxML's GTRGAMMA model with 1,000 bootstrap resamplings. The tree was visualized with Dendroscope and bootstrap values below 50 removed. *E. fergusonii* serves as an outgroup and the corresponding branch is not to scale. Bovine-associated *E. coli* are indicated by colored cows, and both *E. coli* pathotypes and phylogroups are designated with a color code. ST numbers from the MLST analysis for each strain are given in parentheses. *E. coli* isolated from cows are widely distributed in the phylogroups and both commensal and MAEC strains are interspersed in the phylogenetic groups with a polyphyletic history.

To enhance backwards compatibility, we determined the sequence types (ST) for all strains analyzed in the WGA phylogeny according to the Achtman *E. coli* MLST scheme (Additional file 3: Table S3) [13]. The calculated minimum spanning tree (MST) supports the phylogenetic history depicted in the WGA phylogram and confirms the diversity of bovine-associated *E. coli* (Additional file 4: Figure S1). ST10, with nine occurrences, is the most common ST in the 25 *E. coli* genomes from the bovine-associated strain panel. In fact, all bovine-associated *E. coli* of phylogroup A are members of clonal complex 10 (CC10), except for *E. coli* ECA-O157 (ST398, CC398). Nevertheless, the majority of the 25 *E. coli* genomes have different STs, corroborating their phylogenetic diversity.

The polyphyletic evolutionary history of bovine *E. coli* (both MAEC or commensals) is substantiated by their high genotypic and phenotypic plasticity [5,6,9,11,29]. In light of these studies and the genealogy of the bovine-associated *E. coli* in this study (Figure 1) the strain panel is suitable and sufficiently diverse in its phylogeny for more detailed comparative analyses of MAEC and commensal bovine *E. coli* genomes. Two possible explanations for the phylogenetic diversity of MAEC and bovine commensals can be considered. On the one hand, the ability to cause mastitis could have been developed in parallel on several independent occasions during the evolutionary history of *E. coli* by selecting forces [35]. On the other hand, MAEC might be recruited from the normal intestinal commensal microbiota and the ability to cause mastitis is facultative, as has been proposed for ExPEC [14,21,24,25,54].

Gene content correlates with phylogenetic lineages of bovine-associated *E. coli*

The phylogenetic history of *E. coli* can only be consistently replicated using long DNA stretches of the core genome, because otherwise the high rate of homologous and non-homologous recombination in the species obscures phylogeny [13–15,55]. MGEs and HGT introduce DNA regions into the flexible genome, which have their own phylogenetic history that is independent of their host bacteria and can be shared even between distant prokaryotic species [28,56]. However, the core genome is relatively unaffected by such recombination compared to the flexible genome and thus suitable for examining *E. coli* phylogeny [15,21,56]. Several studies showed that recombination between extant *E. coli* phylogroups is limited by phylogenetic diversity, i.e. recombination is higher between closely related phylogroups and within

phylogroups (especially A and B1) [55–57]. This is probably due to the sharing of similar ecological niches and the resulting genomic divergence. Thus, the phylogenetic background of *E. coli* has a big impact on possible recombination events and most importantly on the gene content of the flexible genome [15,21]. Nevertheless, there are examples of convergent evolution in *E. coli,* especially in IPEC pathotypes from multiple parallel phylogenetic origins that typically contain a specific set of VFs, e.g. the occurrence of EHEC in the distant phylogroups B1 and E mediated by HGT of MGEs [46,55].

Despite the phylogenetic diversity of the bovine-associated *E. coli*, we were interested to see if functional convergence of bovine MAEC or commensals exists. There might be a defining subset of genes or VFs for MAEC from different phylogenetic backgrounds that would point to a putative MPEC pathotype. For this purpose we determined the similarity of the genomes based on the presence/absence of all orthologous groups (OG) calculated for the strain panel. Such an analysis, visualized as a so-called gene content tree, has the advantage of considering the core as well as the flexible genome, in contrast to the WGA core genome phylogeny (in which the flexible genome is intentionally filtered out in order to maximize the robustness of the inferred phylogenetic history). Thus, this method can be used to detect functional similarities based on a similar gene content. We clustered all strains based on gene content by calculating the best scoring maximum likelihood (ML) tree of the binary matrix representing the presence and absence of OGs (Additional file 5: Table S4). The topology of the resulting gene content tree mirrors the phylogenetic lineages of the WGA phylogeny with high analogy (Figure 2). All bifurcations that define phylogroups in the gene content tree have high bootstrap values. For comparison purposes we visualized the high similarity between WGA genealogy and gene content tree in a tanglegram (Additional file 4: Figure S2 A and B). This diagram shows that not only the phylogroups are conserved, but also the phylogenetic relationships between individual *E. coli* isolates within the phylogroups. However, some minor differences in the bifurcations between phylogeny and gene content clustering were detected. The two biggest differences concern the placement of phylogroups B2/E and MAEC ECA-O157. In contrast to the WGA-based phylogeny, which clusters phylogroups B2 and E outside the A/B1 sister taxa, the gene content dendrogram places these phylogroups closer to B1 than A. This appears to be due to a more similar gene content, as phylogroups B2/E have a higher recombination frequency with phylogroup B1 than with A [55,56]. Strain ECA-O157 represents an outlier branch in comparison to all other included *E. coli* genomes based on gene content. As this strain is the only strain in phylogroup A that does not belong to the closely related CC10 cluster, this explains its gene content divergence to the other A strains in the gene content tree, which is also apparent in the WGA core genome phylogeny. However, the outlier-clustering of ECA-O157 might also be a result of the high fragmentation of the draft genome and the resulting uncertain accuracy of CDS predictions on which OG analyses are dependent.

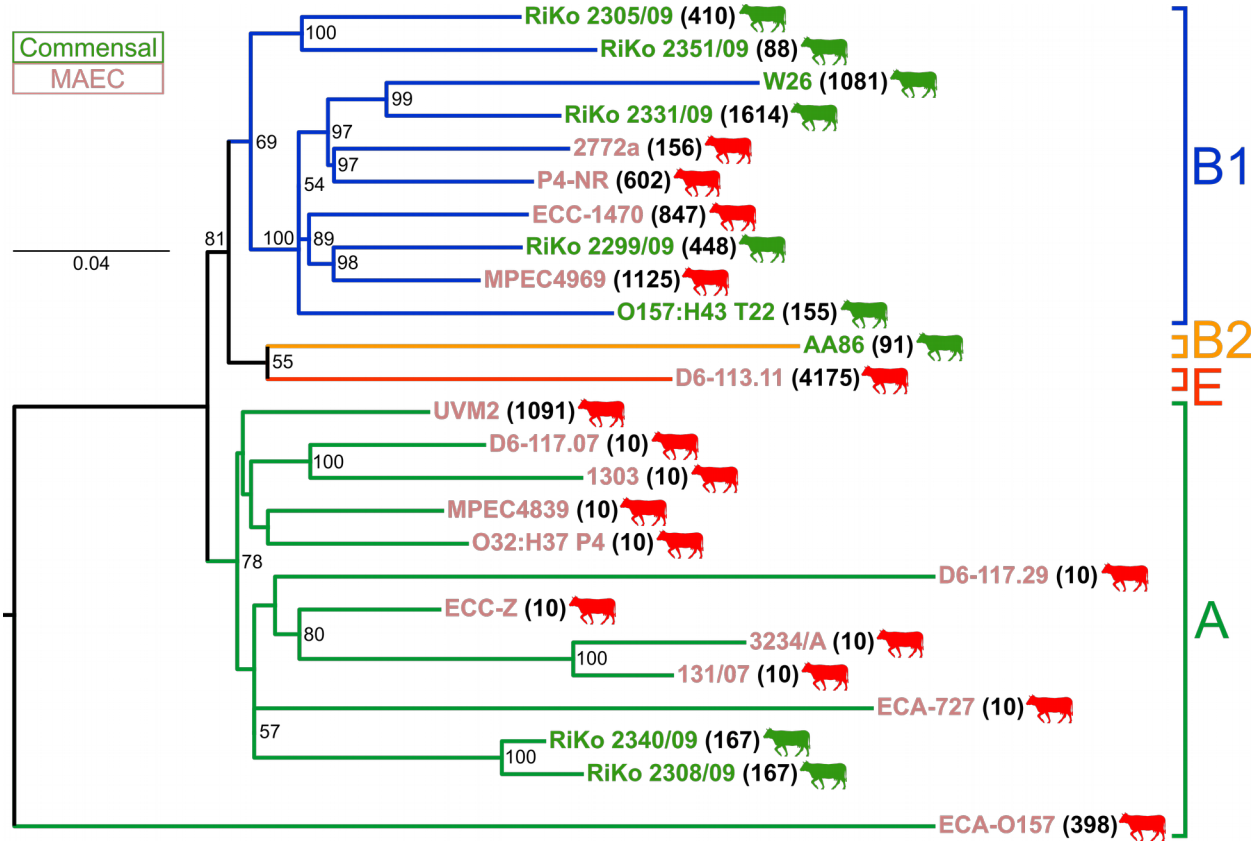

Figure 2 Gene content clustering tree of the bovine-associated *E. coli*. The gene content best scoring ML dendrogram is based upon the presence or absence of orthologous groups (OGs) with 1000 resamplings for bootstrap support values. The tree was visualized midpoint rooted with FigTree and bootstrap values below 50 removed. The distance between the genomes is proportional to the OGs present or absent. The tree topology of the gene content tree follows closely the core genome WGA phylogeny. There is no functional convergence between MAEC or commensal strains, rather a highly diverse gene content.

In conclusion, no functional convergence of bovine MAEC or commensals could be detected and the phylogenetic diversity of the strains is also apparent in a highly diverse gene content. There is no HGT of large genomic regions that are mastitis-specific and the phylogenetic background of the strains has a deciding impact on the overall gene content. A clustering of strains of the same pathotypes would have hinted towards a common gene content and a difference in ecological lifestyles and habitats, as a result of positive selection on the ancestral genomes [55]. However, in this case the flexible genome and the core genome appear to coevolve.

Two previous studies with similar methodology came to two different conclusions. Blum *et al.* [32] reasoned that three mastitis strains (O32:H37 P4, VL2874, VL2732) were much more closely related in gene content compared to an environmental (commensal fecal) strain (K71), based on the different pathotypes. However, the MAEC in this study are phylogenetically strongly related (phylogroup A) whereas the single commensal strain belongs to phylogroup B1.

Thus, as we observed in our study, the phylogenetic relationship had a strong impact on the gene content dendrogram. Kempf *et al.* [34] comparing four phylogroup A MAEC (D6-117.07, O32:H37 P4, VL2874, VL2732), one phylogroup E MAEC (D6-113.11), and the K71 commensal, achieved results comparable to ours. The authors argued that mastitis pathogens with different phylogenetic histories might employ different virulence strategies to cause mastitis, similar to the variable VF repertoire of ExPEC, a hypothesis we tested in this study by searching for well-known *E. coli* VFs in the bovine-associated *E. coli* strains below.

MAEC possess no virulence-attributed orthologs in comparison to commensal strains

It was suggested that the genome content of MAEC is distinct from bovine commensals and not random, as a result of selective pressure. *E. coli* encoding for VFs and FFs important for mastitis pathogenicity are supposedly positively selected within the bovine udder [32,35]. Since no large scale gain or loss of bovine MAEC- or commensal-associated genes could be detected in the gene content tree, we looked into the distribution of OGs, in order to search for genotypic traits enriched in bovine mastitis or commensal isolates. From our point of view, only the comparison of a larger set of MAEC genome sequences with that of bovine commensals is suitable to address this question. If any VFs/FFs existed, that play an important role in the pathogenesis of MAEC, we would expect a wide distribution of the encoding genes among MAEC strains compared to commensals.

The pan-genome of the 25 bovine-associated *E. coli* strains amounted to 116,535 CDS and a total of 13,481 OGs using BLASTP+ with 70% identity and coverage cutoffs. Because of the open nature of the *E. coli* pan-genome [58], all genomes included OGs, which were absent in any other compared strain (so-called singletons; Additional file 6: Dataset S1). The largest numbers of singletons were detected in the highly fragmented genomes of strains D6-117.29 (n=455), ECA-O157 (n=865), and ECA-727 (n=615), a likely consequence of the high number of contig ends and uncertain open reading frame (ORF) predictions. Also, large numbers of singletons in genomes AA86 (n=422) and D6-113.11 (n=361) are to be expected, as these are the only compared genomes of their respective phylogroups, B2 and E. The majority of singletons encode typical proteins of the flexible genome, like hypothetical proteins, proteins associated with MGEs (transposases, phages), restriction modification systems, O-antigen biosynthesis, CRISPR, conjugal transfer systems, and sugar transport/utilization operons. Although several of these genes and gene functions have previously been identified as MAEC-associated in small strain panels [32,34], they most likely play no role in mastitis because of their presence in commensals and/or low prevalence in MAEC.

We searched for OGs that were exclusively present in either the compared MAEC or commensal genomes. However, using these strict initial inclusion and exclusion cutoffs for OGs, we could not detect any genetic traits enriched in either genome group. Hence, we lowered the inclusion and exclusion cutoffs to include OGs present in at least 70% of one pathotype and maximally 30% of the other (see methods section “Ortholog/paralog analysis” for the actual genome numbers). Lowering the thresholds we detected 36 “MAEC-” and 48 “commensal-

enriched" OGs, 84 pathotype-enriched OGs altogether (Figure 3A and Additional file 7: Dataset S2). An "all-strain" soft core genome (as defined by Kaas *et al.* [26]) with this 70% inclusion cutoff (18 genomes of the 25 total) included 3,842 OGs, which is about 82% of the average genome CDS number (Additional file 8: Table S5). Although the size of this soft core genome is relatively large, it is still comparable to prior studies with 3,051 or 3,492 OGs [26,35]. By maintaining a 70% inclusion cutoff, the effects of potential uncertain CDS predictions in the many draft genomes of the strain panel are buffered. Nevertheless, the need for such relaxed inclusion/exclusion cutoffs illustrates the difficulty of detecting any pathotype-enriched OGs at all.

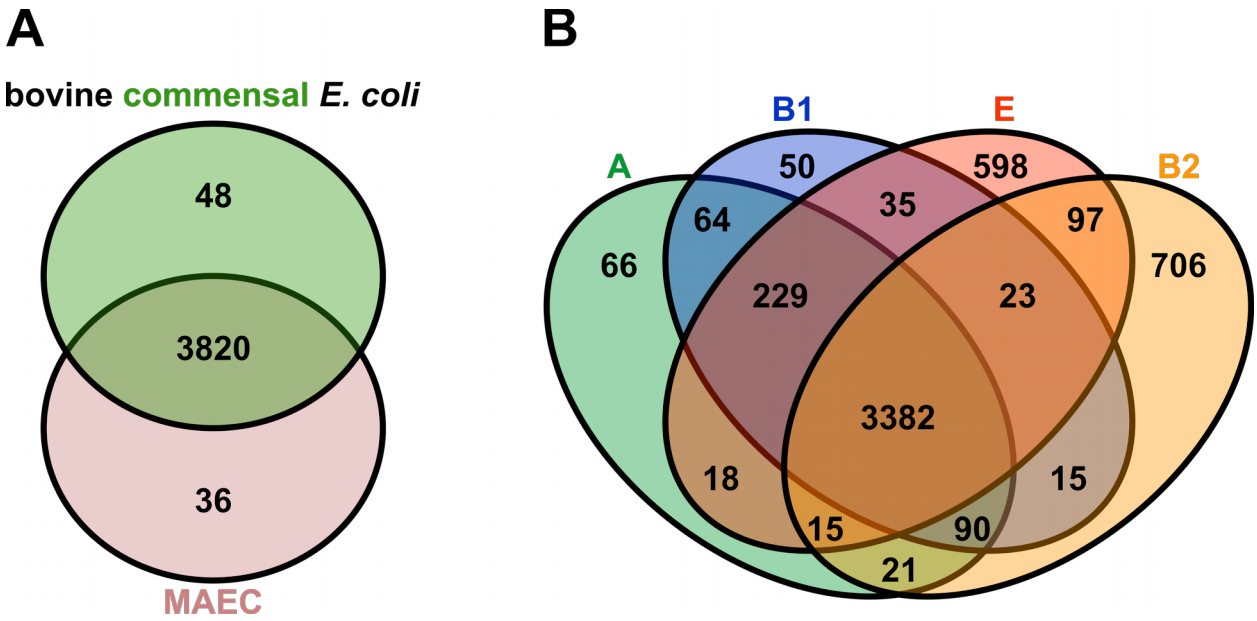

Figure 3 Venn diagrams for gene family enrichment in pathotypes or phylogroups. Enrichment of OGs was determined with 70% inclusion and 30% exclusion group cutoffs by classifying the bovine-associated *E. coli* either in pathotype (MAEC or commensal) or phylogenetic groups (A, B1, B2, or E). Only very few OGs could be detected as pathotype enriched. Instead, OG distribution is strongly affected by phylogenetic background.

Applying said 70%/30% cutoffs to the individual pathotype genome groups separately, instead of all bovine-associated *E. coli* genomes combined, resulted in a "pathogroup" soft core genome of 3,820 OGs (Figure 3A and Additional file 7: Dataset S2). This is different to the "all-strain" soft core genome because of the inclusion of OGs which are present in at least 70% of the genomes of both pathotype groups (instead of 70% of all 25 genomes). Because phylogeny was shown to exhibit a strong effect on the gene content of *E. coli* isolates, we evaluated the detected 84 pathotype-enriched OGs further. For this purpose, we determined if any of the pathotype-enriched OGs can also be classified as phylogroup-enriched OGs. In this case, gene content similarities through shared ancestry might overshadow functional relatedness and thus we defined these pathotype-enriched OGs as "non-significant". We used the same cutoffs to identify OGs that were enriched in genomes of the four phylogroups (A, B1, B2, and E; Figure 3B and Additional file 9: Dataset S3). This analysis resulted in many phylogroup-enriched OGs, supporting the impact of phylogeny on gene content and the resulting observation of the

similarity between the gene content tree and the WGA phylogeny. Analogous to the above observed singleton-content, phylogroups B2 and E had a relatively high number of enriched OGs, because both groups only included one genome each. For our significance evaluation, we compared the proteins of the phylogroup-enriched as well as the all-strain and phylogroup soft core OGs (Additional file 8: Table S5 and Additional file 9: Dataset S3) to the proteins of the pathotype-enriched OGs with BLASTP+ and 70% identity and coverage cutoffs. Of the determined 36 MAEC- and 48 commensal-enriched OGs, only 14 and 13 did not have a phylogroup-enriched counterpart and were marked “significant”, respectively. All others had an analogous hit in phylogroup-enriched categories or a soft core genome (Additional file 7: Dataset S2) and were thus labeled as non-significant. Interestingly, a tendency could be observed that phylogroup A (and, to a lesser extent, E) BLASTP+ hits were more associated with the respective MAEC-enriched OGs whereas phylogroup B1 BLASTP+ hits associated more with commensal-enriched OGs. Although it is tempting to assume putative MAEC VFs are therefore associated with phylogroup A, this result might just corroborate the finding that MAEC are more common in phylogroup A and commensals in phylogroup B1. Overall, the mastitis-enriched OGs are mostly functionally involved in the dissemination (integrase) or maintenance (toxin-antitoxin systems and phage proteins) of MGEs. This might be an indication for different habitats or selection pressures leading to some mobile elements not being exchanged between different phylogroup isolates, respective commensal and MAEC isolates. Nevertheless, this distribution might just be a result of phylogroup A being the predominant lineage for MAEC and B1 for commensals in our strain panel and background noise of overall *E. coli* genome plasticity (Table 1).

Commensal-enriched orthologous groups are associated with fitness factors

The 13 “significant” commensal-enriched OGs, which did not yield a hit in the phylogroup-enriched categories, included two interesting gene clusters (Additional file 7: Dataset S2). The first one is the core LPS biosynthesis cluster between genes *waaP* and *waaC* in the strain RiKo 2299/09. These four proteins (locus tags RIKO2299_91c00430 to RIKO2299_91c00460) are annotated as: WaaO (UDP-glucose:(glucosyl) LPS alpha1,3-glucosyltransferase), WaaT (UDP-galactose:(glucosyl) LPS alpha1,2-galactosyltransferase), a “lipopolysaccharide core biosynthesis protein”, and WaaW (UDP-galactose:(galactosyl) LPS alpha1,2-galactosyltransferase). Although these genes are involved in the core LPS biosynthesis (and not the variable O-antigen), the implied putative conservation is interesting, as all commensals have different serotypes (Table 1). LPS is considered to be a key factor for MAEC in eliciting mastitis. The detection of LPS by host receptors (Toll-like receptor 4, TLR4) triggers the signal cascade that leads to the inflammatory response of the bovine udder [3,7]. Additionally, long-chain LPS might be important during the colonization of the mammary gland, as its expression contributes to serum resistance and virulence [59]. The second gene cluster overrepresented in the bovine commensal *E. coli* isolates tested is the aerobactin siderophore biosynthesis operon (*iucABCD*) with the siderophore receptor-encoding (*iutA*) and the associated putative transport protein ShiF-encoding genes (locus tags in strain RiKo 2340/09 RIKO2340_186c00010 to RIKO2340_186c00060). It is surprising that both of these gene clusters are associated with commensals. Aerobactin is considered to be an ExPEC VF needed for iron uptake under limiting

conditions, e.g. in the urinary tract or serum [60,61]. The remaining four significant enriched OGs in bovine commensal *E. coli* have an IS element (insertion sequence), "putative", or "hypothetical protein" annotation (paralogs RIKO2340_128c00050/RIKO2340_203c00010, RIKO2340_203c00020, and RIKO2340_65c00330, respectively) and unknown functions. The 35 non-significant putative commensal-enriched OGs include fimbrial genes, genes of the lactose/cellobiose (*bcgHI*) and galactitol (*gatZCR*) phosphotransferase systems (PTS), Cas CRISPR genes (*casCD*), a short putative arylsulfatase (53 aa), genes for sucrose catabolism (*cscBKAR*), a putative ABC transporter, a lipoprotein, a type II secretion system (T2SS-2) protein GspG, a non-characterized autotransporter, and a DNA adenine methyltransferase. Their role for bovine commensalism remains unclear, especially because of their additional association with phylogroups.

MAEC-enriched orthologous groups are mostly associated with mobile genetic elements

The 14 “significant” mastitis-enriched OGs do not include any coherent gene cluster (Additional file 7: Dataset S2), but eight are located in close proximity to each other in the genome of strain 1303 (*rzpQ* EC1303_c16730, *ydfR* EC1303_c16750, *quuQ_1* EC1303_c16790, *relE* EC1303_c16830, *relB* EC1303_c16840, *flxA* EC1303_c16860, EC1303_c16890, EC1303_c16900). All of these proteins belong to a prophage without noticeable features (see 1303 prophage2 below). Additionally, four non-significant genes, *ydfO* ("uncharacterized protein", E1303_c16680), *cspI* and *cspB* ("cold shock proteins", EC1303_c16710 and EC1303_c16770), and *recE* ("exonuclease VIII, 5' -> 3' specific dsDNA exonuclease", paralogs EC1303_c12230/EC1303_c16970), also fall into the same prophage region. Because the prophage genome does not contain genes related to metabolic or virulence functions, the role of the encoded gene products in mastitis cannot be determined. The ninth significant OG, *ylbG* (E1470_c05180), is a putative DNA-binding transcriptional regulator. The tenth mastitis-enriched gene, *ybbC* ("putative immunity protein", EC1303_c04920; in strain ECC-1470 E1470_c05150), is just two genes and one pseudogene upstream of *ylbG*. Both proteins are associated with an *rhs* element directly upstream. Because of their repetitive nature, Rhs elements (rearrangement hotspot) have been implicated in promoting recombination. Although the functions of Rhs proteins are poorly understood, they have been shown to mediate intercellular competition similar to contact-dependent growth inhibition systems [62,63]. Intercellular competition and cell-to-cell communication are important features for *E. coli* living in bacterial communities. Still, their function in mastitis pathogenesis is unclear, but might be related to biofilm formation [63]. The eleventh significant OG, *insF1_2* (EC1303_c10450), is part of an IS element within a genomic island of strain 1303 (see GI4 below). EC1303_c10450 is a part of the *pga* gene cluster required for poly-*N*-acetylglucosamine (PGA) synthesis. This extracellular polysaccharide is associated with *E. coli* biofilm formation [64,65] and can contribute to host innate immune evasion and virulence in several bacterial pathogens [66–68]. The downstream adjacent genes of the cluster, *ymdE* and *ycdU*, but not the *pga* genes themselves, were associated with mastitis isolates of phylogroup A [35]. *ymdE* (EC1303_c10470) was classified as a non-significant mastitis-enriched OG in our analysis, because of its affiliation with phylogroups A and E. Based on these results the role of PGA-dependent phenotypes, such as biofilm formation in mastitis pathogenesis, has to be reconsidered. However, a recent review put biofilm formation in context to mastitis

virulence, especially for facilitating persistence in the mammary udder [69]. Protection against the host immune response may generally promote the ability of *E. coli* strains to successfully colonize and propagate in the udder, and cause mastitis. Finally, the antitermination protein-encoding genes *quuD* and *quuQ_2* (EC1303_c12460 and EC1303_c45980, a paralog of *quuQ_1* EC1303_c16790) were located within prophage regions present in the genome of MAEC 1303 (see prophage1 and 4 below, respectively). The last two significant MAEC-enriched OGs are hypothetical proteins and have unknown functions (paralogs EC13107_29c00810 and EC13107_144c00080 of strain 131/07 and paralogs EC2772a_26c01800 and EC2772a_91c00020 of strain 2772a). According to the sequence contexts in *E. coli* 131/07 and 2772a, these genes are most likely located in prophage regions as well.

In summary, the putative mastitis-eliciting function of any of the genes within the MAEC-enriched OGs is unclear. No traditional *E. coli* VFs have been found among MAEC-enriched OGs. However, two non-significant MAEC-enriched OGs are involved with virulence, the *aslA* and *eprI* genes, which were also associated with the phylogroups A/B2/E and A/E, respectively. The arylsulfatase AslA has been implicated in the invasion of the blood-brain barrier by MNEC [70]. Invasion of mammary epithelial cells has been suggested to play a role in persistent *E. coli* mastitis [4,6]. It is, however, not clear whether acute/transient or persistent/chronic bovine mastitis outcomes depend on a specific *E. coli* genotype [32]. The *eprI* gene belongs to the *E. coli* type three secretion system 2 (ETT2) determinant, which is a large gene cluster with frequent deletion isoforms in *E. coli* [71]. ETT2 is also implicated in being involved in invasion and intracellular survival of blood-brain barrier cells of MNEC K1 strains [72], but also in serum resistance even in an incomplete form [61]. Its prevalence has been analyzed in bovine mastitis *E. coli* isolates and determined to be approximately 50%, but mutational attrition was numerous [73]. Thus, a role of ETT2 in MAEC is debatable, especially since only *eprI* and none of the other ETT2 genes were MAEC-enriched.

Genomic islands and prophages in MAEC 1303 and ECC-1470 contain only few well-known virulence-associated genes

Some of the above identified mastitis-enriched genes were associated with prophages or GIs. Bacterial VFs are generally over-represented and accumulated on mobile, or formerly mobile genetic elements, especially pathogenicity islands (PAIs) [74,75]. Because MGEs are prone to contain repetitive sequences, the short sequencing reads of most current high-throughput sequencing technologies cannot be unambiguously assembled in these regions [76]. Additionally, automatic ORF prediction as well as annotation still remains a challenge in MGEs. Thus, we identified prophages and GIs only for the two closed 1303 and ECC-1470 MAEC genomes.

Both *E. coli* 1303 and ECC-1470 genomes include several putative pathogenicity, resistance, and metabolic islands, as well as prophages (Additional file 10: Dataset S4 and Additional file 11: Dataset S5). Many of the GIs and prophages are flanked by tRNAs, which are hotspots for chromosomal insertion of GIs and bacteriophages [28]. GIs could only be detected in the chromosomes of the closed genomes, but not on the respective plasmids. However, on the F

plasmid present in *E. coli* 1303, p1303_109, a smaller 17-kb transposable element was identified. Mastitis isolate 1303 additionally harbors an episomal circularized P1 bacteriophage [77], designated p1303_95.

Generally, the genome of mastitis isolate 1303 includes twelve GIs ranging in size from 11 to 88 kb and encoding from 11 to 81 CDSs (Additional file 10: Dataset S4). One large composite GI (GI4) combines pathogenicity and resistance-related genes. It partly contains the biofilm-associated polysaccharide synthesis *pga* locus plus flanking genes identified as mastitis-associated by Goldstone *et al.* [35]. The resistance-related genes of GI4 are located on the AMR-SSuT island (antimicrobial multidrug resistance to streptomycin, sulfonamide, and tetracycline), which is prevalent in *E. coli* from the bovine habitat [78,79]. The encoded resistance genes are *strAB*, *sul2*, and *tetDCBR*. A comparison of the corresponding genomic region of *E. coli* 1303 with two publicly available AMR-SSuT island sequences from *E. coli* strain SSuT-25 [78] and *E. coli* O157:H7 strain EC20020119 [79] is shown in Additional file 4: Figure S5. Transposon Tn*10*, also present on the resistance plasmid R100, is an integral part of the AMR-SSuT island and comprises the *tetDCBR* genes. This highlights the composite nature of the AMR-SSuT island and of GI4 in general. The resistance markers of AMR-SSuT are prevalent, as seven strains of the panel contain some or all of the genes (D6-117.29, ECA-727, RiKo 2305/09, RiKo 2308/09, RiKo 2340/09, RiKo 2351/09, and W26). The AMR-SSuT island provides additional secondary selective advantages to strains in the gastrointestinal tract of cattle, independent of the antibiotic resistances, which might be one reason for its wide distribution [78].

The twelve GIs harbored by mastitis isolate ECC-1470 vary in size between 8 to 58 kb and code for 9 to 61 CDSs (Additional file 11: Dataset S5). *E. coli* ECC-1470 (Ont:H2) encodes for a flagellin of serogroup H2 (100% identity to *fliC* of EHEC O103:H2 strain 12009, accession-number: BAI30971.1) and an uncharacterized small alternative flagellin, FlkA, encoded on GI10. The neighbouring *flkB* gene encodes for a FliC repressor. This small alternative flagellin islet can elicit unilateral H-antigen phase variation [80–82]. The MAEC strain P4-NR (O15:H21/H54), which usually expresses a serotype H21 flagellin (100% identity to *fliC* of an O113:H21 EHEC, accession-number: ABI23966), also harbours a similar alternative flagellin system determinant consisting of the serotype H54 flagellin gene *flmA54* (97% identity to *flmA54* of *E. coli* E223-69, accession-number: BAD14977.1) and the associated *fliC* repressor-encoding gene *fljA54* [83]. Phase variation is an important mechanism for bacterial adaptation to different environments and especially for pathogens to evade the host's immune system [80]. GI12 of ECC-1470 is a large PAI containing a fimbrial operon of the P adhesin family (*pixGFJDCHAB*, *pixD* is a pseudogene), a phosphoglycerate transport operon (*pgtABCP*), the putative MAEC-associated Fec transport operon (*fecEDCBARI*), the 9-O-acetyl-*N*-acetylneuraminic acid utilization operon (*nanSMC*), and the type 1 fimbriae operon (*fimBEAICDFGH*). This PAI is a composite island with the 5'-end similar to PAI V from UPEC strain 536 with the *pix* and *pgt* loci, also present in human commensal *E. coli* A0 34/86, [84–86] and the 3'-end similar to GI12 of MAEC 1303 with the *nan* and *fim* gene clusters. However, GI12 of strain ECC-1470 lacks the K15 capsule operon present in PAI V of UPEC 536 and has a much smaller size [85]. GI4 codes for a lactose/cellobiose PTS system (*bcgAHIFER*, *bcgI* is a pseudogene). Interestingly, the *bcgHI*

genes were also identified as putative commensal-enriched OGs in our ortholog analysis, but declared as "non-significant" because of their association with phylogroup B1.

Four prophages were predicted in the genome of MAEC 1303 ranging from 29 to 48 kb encoding for 44 to 59 CDSs (Additional file 10: Dataset S4). The enclosed prophage genomes do not comprise many virulence-associated genes, and mostly code for functions required for maintenance and mobilization. The only exception is *bor*, a gene of phage lambda widely conserved in *E. coli* and encoded by strain 1303 chromosomal prophage1. The outer membrane lipoprotein Bor is homologous to Iss (increased serum survival) and involved in serum resistance of ExPEC [87,88]. The lack of putative *E. coli* VFs encoded by prophages is also true for the five predicted prophage genomes of MAEC ECC-1470 (Additional file 11: Dataset S5). Two outer membrane proteins (OMPs) are encoded by ECC-1470 prophage1, the porin NmpC and the omptin OmpT. OmpT is an outer membrane protease and, like Bor, also involved in serum resistance by cleaving antimicrobial peptides [89–91].

In summary, the MGEs of MAEC strains 1303 and ECC-1470 do not carry many known virulence-associated genes, which entail an advantage to mastitis pathogens. Only very few PAI-encoded VFs were detected as pathotype-enriched OGs. It is more likely, that the encoded VFs play a role for commensal colonization or persistence in the bovine gastrointestinal tract. But, the number and variability of GIs and prophages overall support the genomic diversity of bovine *E. coli*, and especially MAEC, and their potential for adaptation to specific growth conditions by gaining or losing large genomic regions [28]. To illustrate the resulting mosaic-like structure of *E. coli*, we created circular genome diagrams for all MAEC 1303 and ECC-1470 replicons indicating the core and the flexible genome by labeling the predicted GIs and prophages (Additional file 4: Figure S3 A and B). Importantly, the MGEs of the closed genomes were not specifically present (or absent) in regard to pathotype, bovine commensals or MAEC.

Virulence or fitness factors present in bovine commensal *E. coli* or MAEC

Several virulence-associated properties have been proposed for MAEC pathogenicity [3,34,92]: multiplication and persistence in milk and the udder [10,93,94], resistance to serum components and neutrophil neutralization mechanisms [7,95,96], adhesion to (and invasion of) mammary epithelial cells [4,6,10,97], and stimulation of the innate immune response by PAMPs [98,99]. These properties are presumably based upon virulence-associated traits, like adherence to host cells, iron uptake, expression of toxins, and factors protecting against the host's innate and adaptive immune response. We detected some of these classical *E. coli* virulence-associated factors in genomes of MGEs of MAEC isolates 1303 and ECC-1470. To examine the distribution of virulence-associated factors in more detail we searched for well-known *E. coli* VFs encoded by the bovine-associated *E. coli* genomes (Additional file 12: Table S6) [100].

Only about half of the 1069 gene products involved in the biosynthesis and function of 200 *E. coli* virulence and fitness-associated factors yielded BLASTP+ hits in the 25 bovine-associated *E. coli* genomes. Virulence-associated proteins of the VF panel present in (556) and absent

from (513) these *E. coli* genomes are listed in Additional file 12: Table S6 (column "present_in_strain_panel", a '1' for presence and a '0' for absence). Results of the BLASTP+ hits for the virulence-associated proteins are listed in Additional file 13: Table S7. Many classical IPEC VFs were not present in the bovine-associated strains [22], of which the most prominent are: dispersin and aggregative adherence fimbriae (EAEC), coli surface antigen and pig-associated K88 pili (ETEC), type 4 pili like EPEC *bfp* bundle-forming pili, and locus of enterocyte effacement (LEE) encoded type III secretion system (T3SS) plus effectors, Shiga toxins, EHS (subtype i1) T6SS, and EHEC alpha-hemolysin. Interestingly, all major virulence factors of EHEC are missing. Furthermore, several VFs associated with ExPEC [20] were absent, such as several typical serine protease autotransporters of *Enterobacteriaceae* (SPATE) like Sat and Pic (type V secretion systems, T5SS), S fimbriae, salmochelin siderophore, colicin V, and colibactin.

The fecal isolate RiKo 2351/09 of phylogroup B1 yielded the most virulence-associated protein hits (297), whereas MAEC ECA-O157 of phylogroup A the fewest (162). There were 241 virulence-associated protein hits on average in the strains included in this study. The few hits yielded by the MAEC ECA-O157 genome sequence might be due to its high fragmentation with 1,173 contigs and putative CDS frameshifts, caused by its low 11-fold 454 sequencing coverage. Nevertheless, overall VF hits did not relate to contig number. We could not detect a correlation between the number of virulence-associated genes and pathotype as both, commensal strains and MAEC, exhibited comparable average virulence-associated genes hits (250 and 237, respectively). The average number of virulence-associated genes was in the same range in the *E. coli* genomes of the different phylogroups (phylogroup A: 233, B1: 254, B2: 227, and E: 235).

We converted the BLASTP+ VF hits for each strain into a presence/absence binary matrix (Additional file 14: Dataset S6) to enable grouping of the compared strains according to their VF content (Additional file 4: Figure S2 C). Most of the genomes belonging to the same phylogroup clustered together. However, there is less agreement with the gold standard WGA phylogeny. Consequently, the association of the strains with phylogroups in the VF content tree is not as well conserved as in the overall gene content tree, as shown by a tanglegram with the WGA phylogeny (Additional file 4: Figure S2 D). HGT as well as homologous and non-homologous recombination of virulence-associated genes are events that confound the determination of the strains' phylogenetic history. For example, phylogroup A strains 1303 and D6-117.29 cluster with B1 strains as a result of the presence of the large Flag-2 MGE (Figure 4). Nevertheless, the phylogenetic background still dominates the clustering without excessive genetic admixture from other phylogroups. Also, based on this clustering, no association with pathotypes was detected. This is in accordance with a previous study analyzing candidate virulence genes in MAEC [34].

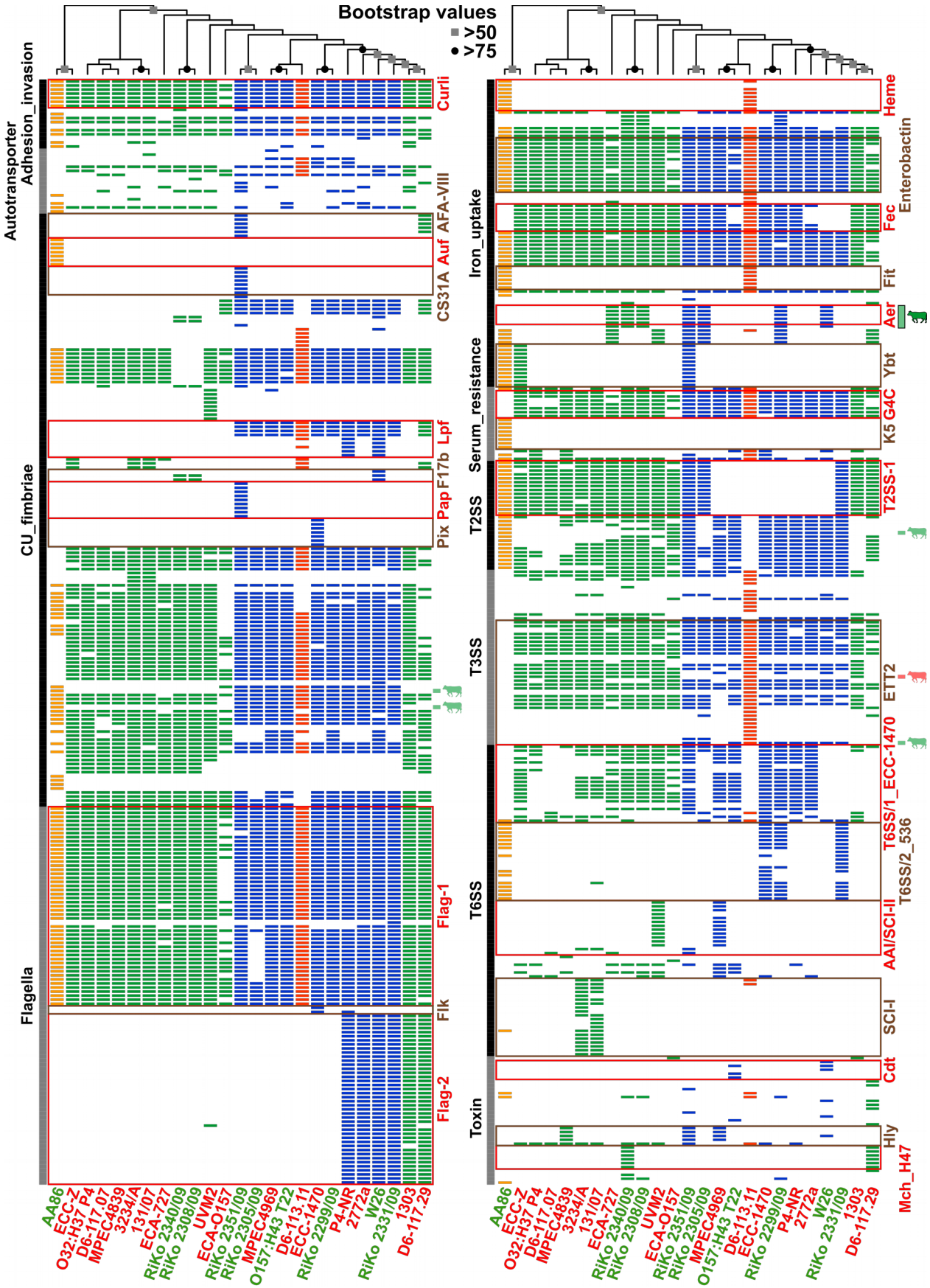

Figure 4 Heatmap indicating presence or absence of virulence factors. Each row of the binary matrix indicates the presence or absence of a virulence-associated gene (a BLASTP+ hit). VF classes are indicated at the side in black and grey. Strain names are color-coded for MAEC (red) or commensal (green) pathotype affiliation, columns for strain phylogroup affiliation (green: A, blue: B1, orange: B2, red: E). The clustering dendrogram attached to the heatmaps is based upon the whole binary dataset (not for each heatmap separately) of a best scoring ML tree with 1000 bootstrap resamplings (a more detailed representation of the cladogram can be found in Additional File 4: Figure S2 C). Bootstrap support values are arbitrarily indicated at the bifurcations of the cladogram. Pathotype-enriched VF genes are indicated for MAEC and commensal isolates by cows in red or green, respectively. Only the aerobactin biosynthesis cluster (Aer) plus transport protein ShiF is commensal-enriched and not associated with a phylogroup (indicated by a black-rimmed and opaque green cow). All other pathotype-enriched virulence-associated genes are designated “non-significant” with a phylogroup association. The genes of well-known and important *E. coli* VFs are highlighted in alternating red and brown squares: Curli = curli fibres, AFA-VIII = aggregative adherence fimbriae AFA-VIII, Auf = fimbrial adhesin, CS31A = CS31A capsule-like antigen (K88 family adhesin), Lpf = long polar fimbriae, F17b = F17b fimbriae, Pap = P/Pap pilus, Pix = Pix fimbriae, Flag-1 = *E. coli* peritrichous flagella 1 gene cluster, Flk = alternative *flk* flagellin islet, Flag-2 = *E. coli* lateral flagella 2 gene cluster, Heme = *chu* heme transport system, Enterobactin = enterobactin biosynthesis/transport gene cluster, Fec = ferric iron(III)-dicitrate uptake system, Fit = ferrichrome iron transport system, Aer = aerobactin biosynthesis cluster with *iutA* receptor, Ybt = yersiniabactin iron transport system, G4C = group 4 capsule, K5 = K5 capsule, T2SS-1 = *gsp* general secretion pathway 1, ETT2 = *E. coli* type three secretion system 2, T6SS/1_ECC-1470 = MAEC ECC-1470 subtype i1 T6SS/1, T6SS/2_536 = UPEC 536 subtype i2 T6SS 2, AAI/SCI-II = EAEC 042 subtype i4b T6SS 3, SCI-I = EAEC 042 subtype i2 T6SS 2, Cdt = cytolethal distending toxins, Hly = alpha-hemolysin, Mch_H47 = microcin H47. Clustering of the strains according to virulence-associated gene presence/absence also follows mostly the phylogenetic history of the strains, no clustering of pathotypes was detected. Both MAEC and commensal isolates are distinguished by the lack of classical pathogenic *E. coli* VFs. The same heatmap, but including gene names/locus tags, can be found in Additional File 4: Figure S4.

The presence and absence of the VFs in the different strains were visualized in a heatmap in which the respective genome columns are ordered according to the clustering results (Figure 4). The heatmap is replicated with the corresponding virulence-associated gene names in Additional file 4: Figure S4. The suboptimal CDS prediction in the fragmented ECA-O157 genome is apparent in several hit gaps in the heatmap for putative VFs which are generally present in the strain panel.

Analogous to the all-strain soft core genome we determined an “all-strain” soft core VF set. Against the background that many fragmented draft genomes are included in the strain panel, we once more applied a 70% inclusion cutoff. As a result, virulence-associated genes were included if they were present in at least 18 of the 25 bovine *E. coli* genomes analyzed. The resulting 182 virulence-associated genes (Additional file 15: Table S8) included determinants generally considered to be widely present in *E. coli* isolates, like the Flag-1 flagella system, the

operons encoding type 1 fimbriae, and the *E. coli* common pilus (ECP). But also curli fimbriae, the lipoprotein NlpI, outer membrane protein OmpA, and several iron transport systems (ferrous low pH (*efe*/*ycd*), enterobactin (*ent*, *fes*, and *fep*), ferrous (*feo*), and ferrichrome (*fhu*)) are included. Additionally, several T2SS genes, 16 of the 32 ETT2 genes, and two genes from the ECC-1470 T6SS/1, *impA* and a gene coding for a Hcp T6SS effector-like protein (E1470_c02180), are enclosed. Interestingly, the ferric iron(III)-dicitrate uptake determinant is contained in this soft core VF set. Scavenging iron is an important trait for mastitis pathogens, as the host sequesters iron by binding to high affinity molecules in the udder, like lactoferrin and citrate, and thus limiting bacterial growth by “nutritional immunity” [7,101,102]. Because citrate is the main iron-chelating mechanism found in milk during lactation, the Fec system was hypothesized to be the only essential iron transport system for MAEC growth in milk [32,34,35]. However, considering its ubiquitous presence in our strain panel with MAEC and commensals, a preferentially virulence-associated function of the Fec system in mastitis cannot be confirmed.

Some results of the VF gene analysis are especially worth noting, because of their history of being putatively associated with MAEC. For the genomes of *E. coli* strains, 1303, ECC-1470, and AA86, plasmids have been resolved. Five virulence-associated genes are contained on these plasmids as shown in the BLASTP+ hit table (Additional file 13: Table S7). Three of these, *traJ*, *traT*, and *ompP*, have implications for invasion and serum resistance. All three are encoded by the F-plasmids of MAEC 1303 and ECC-1470, p1303_109 and pECC-1470_100. TraJ is an invasion factor of K1 MNEC important for the traversal of the blood-brain barrier [103]. Both TraT and the omptin OmpP are outer membrane proteins associated with serum resistance in *E. coli*. TraT is a lipoprotein, which is present commonly in MAEC, but did not confer serum resistance properties to MAEC in earlier studies [87,104]. OmpP is a homolog of the aforementioned OmpT [90]. *traT* and *traJ* are also encoded on the smaller 65-kb pAA86S plasmid of strain AA86. *ompP* is missing in commensal *E. coli* AA86. Both, *traJ* and *ompP*, are also present in genomes of several other MAEC and commensals. *traT* is nearly universally present and a member of the all-strain soft core VF set. This indicates that F-plasmids are common in *E. coli* within the bovine habitat, but not associated with MAEC (Additional file 4: Figure S3 A and B).

Previous studies have shown a wide spectrum of serum resistance incidence among MAEC, but also commensal bovine *E. coli* [10,11,31,96]. OMPs, such as Bor/Iss, TraT, and omptins, are important in the defense against antibodies, complement system, and defensins in milk [88,90,104]. Although serum resistance is considered to be one of the most important traits of MAEC, the role of these OMPs in bovine *E. coli* mastitis pathogenesis is unclear [10,11,87,104]. Polymorphonuclear neutrophils (PMNs) play a prominent role in the innate host defense. PMNs are recruited to the inflammation site by cytokines, primarily secreted from alveolar macrophages and epithelial cells, to kill the infectious agent by phagocytosis or neutrophil extracellular traps (NET) and strengthen the inflammatory response [7,105]. Thus, NET and phagocytosis evasion was proposed as a virulence factor for MAEC [3]. Capsules are important structures to avoid neutrophil killing by masking surface structures on the bacterial cell and decreasing opsonization with antibodies or complement [61], but neither the group 4 capsule (G4C) nor a member of group 2 capsules, the K5 capsule, showed any association with MAEC.

In fact, the K5 capsule is only present in the B2 phylogroup commensal strain AA86 (*kpsCDEFMSTU-II*). In contrast, the G4C determinant is generally present in the strain panel and included in the all-strain soft core VF set. The ETT2 type III secretion system is contained on 1303 GI8 and ECC-1470 GI9 (Additional file 10: Dataset S4 and Additional file 11: Dataset S5) and has not only been discussed as a VF during mastitis [73], but has also been implicated in serum resistance of APEC O78:H19 strain 789, besides its degenerate form in the strain [61]. ETT2 has different mutational attrition isoforms in the bovine-associated strain panel, supporting the results of an earlier study [73]. But again, overall ETT2 presence was not related to MAEC. Based on the comparative analysis, and in accordance with Blum *et al.* [10] these results suggest that serum resistance is not an essential trait for the ability of MAEC to cause intramammary infections.

An alternative flagellar system (Flag-2) is encoded on 1303 GI1 [106]. The Flag-2 locus encodes also for a type III secretion system in addition to the alternative flagellar system, which might be in cross-talk with ETT2. In contrast to the typical *E. coli* peritrichous flagella 1 gene cluster (Flag-1), which is a polar system for swimming in the liquid phase, the lateral Flag-2 most likely has its functionality in swarming ability over solid surfaces [106]. Flagella are important for motility, but also for adherence during host colonization and biofilm formation [19]. Additionally, flagella might play an important role in the udder for dissemination from the teat and counteracting washing out during milking [10,101]. Bacterial motility is also co-regulated with the upregulation of iron metabolism genes in MAEC [101].

MAEC ECC-1470 also carries two T6SS determinants located on GI1 (designated as the first ECC-1470 T6SS, T6SS/1) and on GI8 (ECC-1470 T6SS/2), respectively. *E. coli* ECC-1470 T6SS/1 is classified as subtype i1 [107] or the second *E. coli* T6SS-2 phylogenetic cluster [108] and T6SS/2 as subtype i2 [107] or the first *E. coli* T6SS-1 cluster [108]. These sophisticated molecular machineries are important for the export of proteins over the cell membranes. Subtype i1 T6SSs generally participate in interbacterial competition, subtype i2 T6SSs target eukaryotic cells and play a role in the infection process of pathogens. All T6SS are implicated in mediating adherence and biofilm formation [108]. The GI1-encoded T6SS/1 was consistently present in strains ECA-O157, ECA-727, and ECC-Z, but only sporadically in human reference commensal strains, and thus associated with MAEC in a preceding study [33]. Nevertheless, the corresponding phenotypes of these systems are mainly unknown and their function, especially any putative role in mastitis, might well be indirect [108].

In conclusion, MAEC are characterized by a lack of “bona fide” VFs [11,29,34]. Instead, the VF variety observed rather mirrors the high diversity of bovine-associated *E. coli*. Although many of these putative VFs are not connected to mastitis virulence, they are still maintained in the genomes. This suggests that they serve as FFs for gastrointestinal colonization and propagation, rather than VFs.

Specific virulence or fitness genes cannot be unambiguously detected for MAEC or commensal bovine isolates

A myriad of previous publications have tried to identify VFs specific for MAEC with varying degrees of success [6,11,29–31,49,50,96,109,110]. However, the results of these studies do not agree upon the identified VFs, which is due to the diversity of MAEC and general bovine *E. coli*. The aforementioned publications follow a classical diagnostic typing procedure by using PCR assays for virulence-associated gene detection. Only Kempf and colleagues applied a bioinformatic approach similar to ours with a candidate VF panel of 302 genes [34]. Our larger strain and selected VF panels enabled a more detailed analysis.

We allocated the analyzed strains into pathotypes (MAEC or commensal) and used the 70%/30% inclusion/exclusion cutoffs to detect VF association with either group, as in the bottom-up ortholog analysis. Uncertain ORF predictions in the draft genomes are buffered by these low cutoffs. Similar to the OG clustering results more stringent thresholds did not yield any pathotype-enriched VFs. With the 70%/30% cutoffs we recovered only one MAEC- and ten commensal-enriched VF genes (Figure 5A, Table 2, and Additional file 16: Dataset S7). As VFs are defined to function in pathogenicity, virulence-related genes associated with commensals are more likely FFs. We also evaluated these pathotype-enriched virulence-associated genes by comparisons to the corresponding phylogroup-enriched virulence-associated genes. As expected from the OG analysis, the phylogroup had a strong impact on VF enrichment (Figure 5B and Additional file 17: Dataset S8). This approach allowed to analyze VFs which were previously correlated with mastitis pathogenicity. One example is the serum resistance-associated gene *bor*, which has no pathotype association. It is rather underrepresented in phylogroup B1 strains, therefore an association with MAEC strains is misleading.

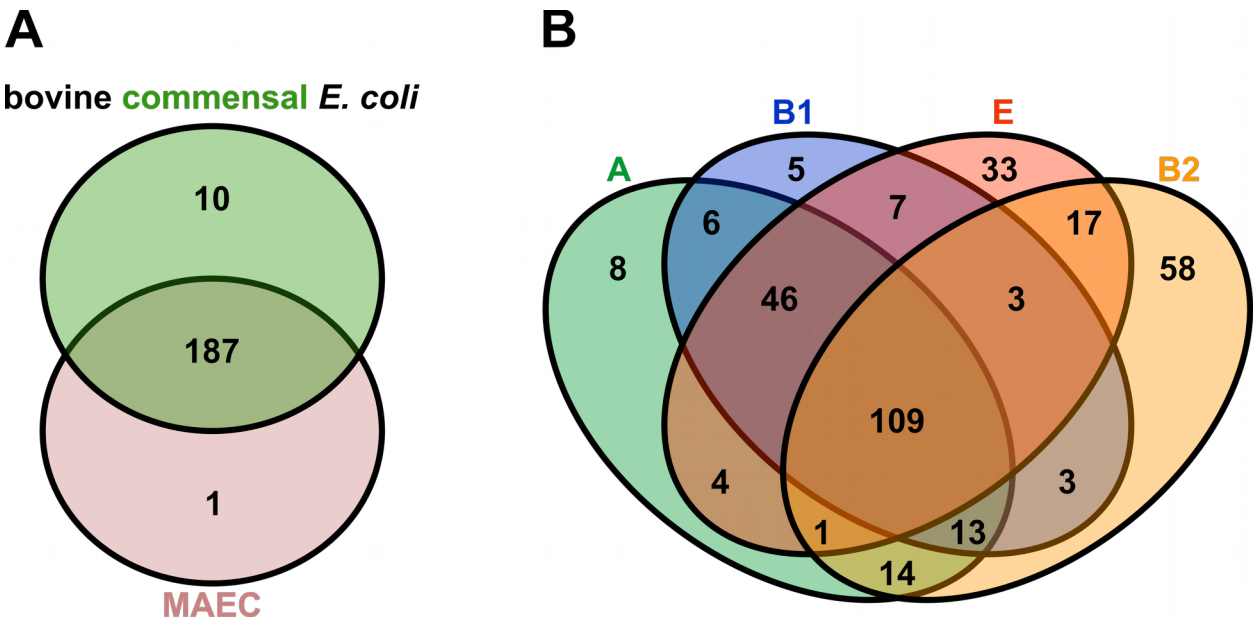

Figure 5 Venn diagrams of virulence-associated gene enrichment in pathotypes and phylogroups. Enriched virulence-associated genes were identified with 70% inclusion and 30% exclusion group cutoffs for the bovine-associated *E. coli* classified either by pathotype (MAEC or commensal) or phylogenetic groups (A, B1, B2, or E). As with the OG enrichment analysis, phylogenetic lineage of the strains dominates VF content and only very few virulence-associated genes were enriched in the pathotypes.

A comparison of the phylogroup-enriched and all-strain/phylogroup soft core VF set to the pathotype-enriched ones identified six “significant” virulence-associated genes not associated with a phylogroup (Table 2). These six commensal-enriched virulence-associated genes are involved in the biosynthesis and transport of the aerobactin siderophore (*iucABCD*, *iutA*, *shiF*). The aerobactin gene cluster was also detected as a significant commensal-enriched OG. Accordingly, this iron uptake system is not an MAEC-specific VF. The aerobactin cluster is often encoded by plasmids harboring additional traits, like colicins and other iron transport systems, e.g. in APEC colicin plasmids [61,111,112]. Thus, its distribution might also be due to positive selection of beneficial traits for commensalism, which are encoded by the same plasmid. The four “non-significant” commensal-enriched VF genes with phylogroup associations represent a part of Yde fimbriae (EC042_1639 and *ydeT*; enriched in phylogroups B1, B2, and E), of the T2SS-2 system (*gspG*, also detected in the ortholog analysis; enriched in B1 and B2), and of the ETT2 system (EC042_3075; enriched in B1 and E). The only mastitis-enriched virulence-associated gene, categorized as non-significant, was *eprI*, which codes for an inner ring component of the ETT2 system. This gene was also identified as a non-significant MAEC-enriched OG with phylogroup A and E association. It is interesting that both MAEC- and commensal-enriched virulence-associated genes are associated with the same ETT2 gene cluster. But, this branched enrichment is not that surprising, because first, ETT2 is a gene cluster with many different mutational attrition isoforms with unknown function in *E. coli* and in particular in MAEC [71,73] and as a consequence not the whole gene cluster was identified. Second, both virulence-associated genes are rather dependent on the phylogenetic background

(associated with phylogroups A/E and B1/E, respectively). This highlights the difficulty in determining a gene as a putative VF for MAEC, especially considering that also half of the ETT2 genes actually belong to the all-strain soft core VF set.

Table 2 Virulence-/fitness-associated genes enriched in MAEC and commensal isolates. Of the eleven genes only the aerobactin gene cluster (*iutA*, *iucDCBA*, *shiF*) enriched in the commensals was not associated with a phylogenetic background. All other genes had also a phylogenetic association, and thus were labeled “non-significant”.

Gene/locus tag	Accession number	VF class	Phylogroup category hit
MAEC-enriched VF gene			
eprI	YP_006097353	T3SS/ETT2	A/E-enriched
Commensal-enriched FF genes			
EC042_1639	YP_006095949	CU fimbriae	A-absent
ydeT	YP_006095947	CU fimbriae	A-absent
iucA	NP_755502	Iron uptake	no hit
iucB	NP_755501	Iron uptake	no hit
iucC	NP_755500	Iron uptake	no hit
iucD	NP_755499	Iron uptake	no hit
iutA	NP_755498	Iron uptake	no hit
shiF	NP_755503	Iron uptake	no hit
gspG	YP_006125848	T2SS-2	B1/B2-enriched
EC042_3075	YP_006097370	T3SS/ETT2	B1/E-enriched

Because ETT2 genes were present in MAEC and commensals, we wanted to analyze the ETT2 determinant in more detail in our strain panel. In addition to ETT2, we also examined the large ECC-1470 T6SS/1 and Flag-2 gene regions. All three putative virulence regions show a high amount of mutational isoforms and/or absence in the strain panel (Figure 4), warranting a detailed analysis. For this purpose, the gene composition of such regions was depicted for all bovine-associated *E. coli* from the strain panel (Figure 6 and Additional file 4: Figure S6 A and B). In the case of strain D6-117.29 the ETT2 and T6SS regions could probably not be fully manually assembled, because of the high fragmentation of the genome.

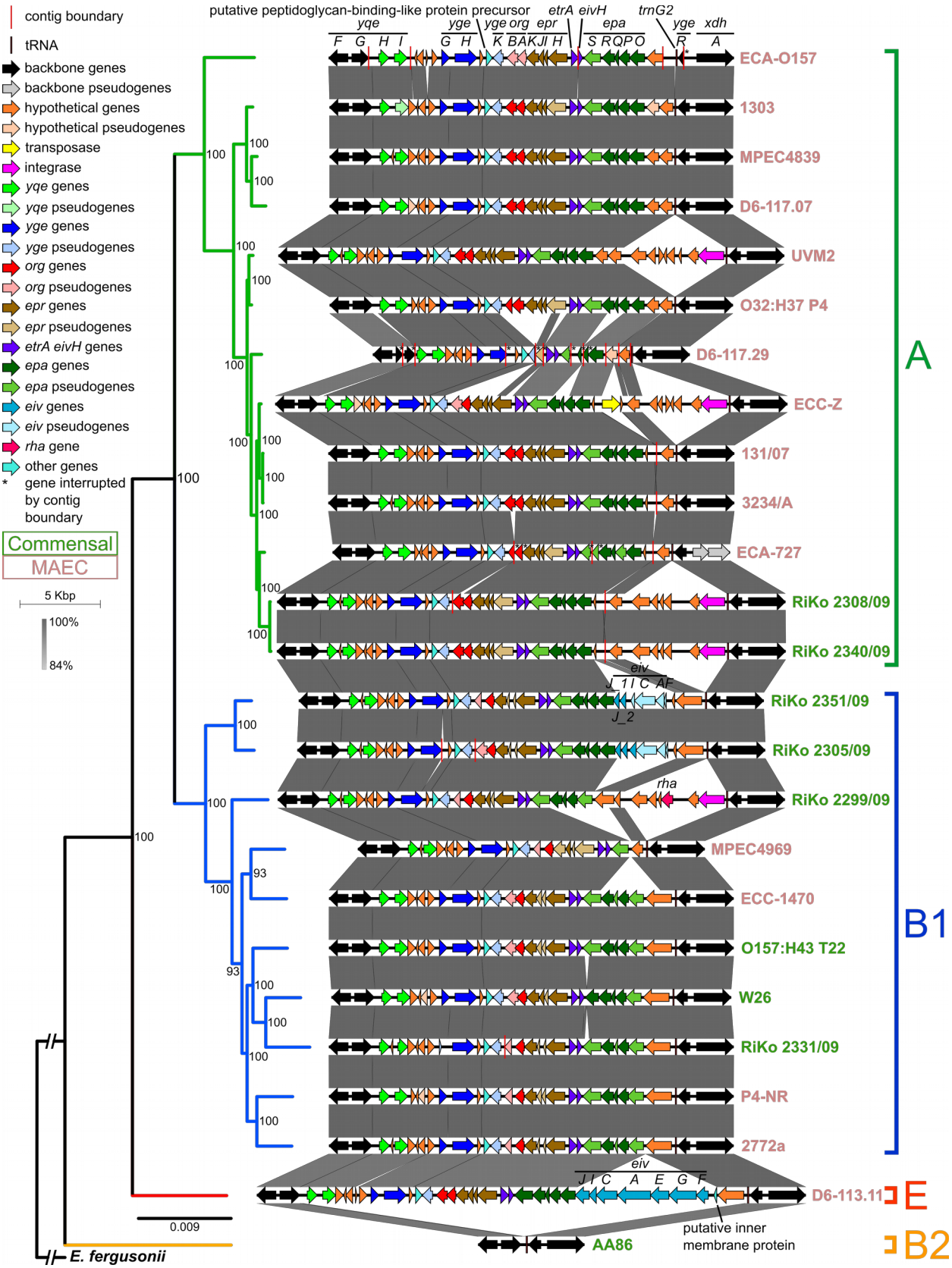

Figure 6 Gene organization of the ETT2 gene cluster in the bovine-associated *E. coli* genomes. Comparison of the ETT2 gene cluster in the *E. coli* of the strain panel based on BLASTN+. Homologous regions are connected via grey vertices and colored by nucleotide identity. The genomes are ordered according to the WGA core genome phylogeny (Additional file 4: Figure S2 A), which is attached on the left side (bootstrap support values below 50 were removed). Phylogroups are indicated correspondingly. MAEC strain names are colored in light red and commensals in green. Gene names are indicated above genomes encoding for these. The respective contigs of the draft genomes containing the gene cluster were concatenated (contig boundaries are indicated by red vertical lines) and CDS spanning contig borders reannotated if needed (indicated by asterisks). ETT2 contigs of genome D6-117.29 were difficult to concatenate, because of its high fragmentation. Backbone genes not belonging to ETT2 are colored black. Genes within the ETT2 region have different colors (see the legend) to be able to evaluate their presence. Pseudogenes have a lighter color fill. ETT2 shows a large number of different mutational isoforms. Nevertheless, ETT2 composition follows phylogenetic history rather than pathotype affiliation.

The ETT2 gene cluster shows high genetic flexibility and many deletions and insertions (Figure 6). Nevertheless, small features still reveal a phylogenetic relationship of similar pseudogene composition. For example *eprJI*, *orgB*, and *epaO* are mostly pseudogenes in B2 strains, but the genes seem to be functional in all phylogroup A and E strains. No comparable pattern was found in relation to the pathotypes. Almost all genomes lack a fragment present in the putatively intact ETT2 region of phylogroup D1 EAEC strain 042 (*eivJICAEGF*), which is located between two small direct repeats and thus often deleted [71]. Only the ETT2 gene cluster in phylogroup E isolate D6-113.11 has an identical structure as 042 (phylogroup E is most closely related to phylogroup D1).

The Flag-2 region is basically present or entirely absent (with two pseudogene remnants [106]) in the strain panel. No intermediate attrition isoforms are observable (Additional file 4: Figure S6 A). The distribution of the Flag-2 gene cluster in different phylogenetic backgrounds suggests three independent insertions of the Flag-2 MGE in the strain panel. Because of the close phylogenetic relationship of the commensals RiKo 2331/09 and W26, and the MAEC strains 2772a and P4-NR, a Flag-2 insertion was most likely in a common ancestor of these strains with vertical propagation. However, a subsequent large deletion is apparent in O157:H43 strain T22 of the same clade. This deletion encompasses the whole Flag-2 region and respective flanking backbone genes. Thus, *E. coli* O157:H43 strain T22 was omitted from the diagram. Additionally, the deletion includes also the housekeeping genes downstream of the T6SS/1 gene cluster of *E. coli* ECC-1470 indicated by dots in Additional file 4: Figure S6 B.

The subtype i1 T6SS/1 of MAEC strain ECC-1470 is the most variable of the virulence-associated regions investigated in more detail in this study, with many repetitive sequence subregions. This further complicates the assembly of this region in the draft genomes. Typical for T6SSs, it is also adjacent to an *rhs* element, which is a highly repetitive chromosomal region [38]. Strain ECA-727 lacks the *yafT* to *impA* genes, because of a putative phage insertion in this region. This phage is not included in the figure and the truncation is indicated by dots in the

diagram. The T6SS determinants in MAEC strains 1303, MPEC4839, D6-117.29, D6-113.11, and commensal RiKo 2305/09 are most likely not functional because of their small sizes. Overall, we could not find any features of this gene cluster, which are associated with phylogeny or pathogenicity of the strains. Thus, the T6SS determinant seems to be a region with a high potential for horizontal gene transfer or rearrangement. Because of a low prevalence of T6SS genes in the five included MAEC genomes and presence in commensal strain K71, another previous study also questioned the role of T6SS systems in mastitis [34]. In conclusion, even our detailed analysis of the ETT2, Flag-2, and T6SS/1 regions did not reveal any association with MAEC isolates in our strain panel. These regions of the flexible genome mirror the underlying genomic and phylogenetic diversity of bovine *E. coli*.

In the ortholog cluster analysis above, we identified a MAEC-enriched gene (EC1303_c10450) which is a part of an IS element. This gene is located within one of the three phylogroup A MAEC-enriched regions described by Goldstone and colleagues [35]. The corresponding region is the biofilm-associated polysaccharide synthesis *pga* locus with adjacent genes, which is also part of GI4 of the MAEC strain 1303. These authors also defined a second genomic region, the *paa* phenylacetic acid degradation pathway plus neighbouring genes (17 genes, *feaRB-tynA-paaZABCDEFGHIJKXY*), as phylogroup A MAEC-associated. As with *paa* only some of the genes were identified as belonging to the mastitis-specific core genome of the study (*feaRB-paaFGHIJKXY*). The third region codes for the Fec uptake system. The *fec* genes are included in the all-strain soft core VF set of our VF panel analysis (Additional file 15: Table S8), thus not associated with pathotypes or phylogroups. Because Goldstone *et al.* included only phylogroup A MAEC genomes into their study (66 in total), we analyzed the *pga* and the *paa* gene regions in detail and applied the BLASTP+ workflow and conversion to a binary presence/absence matrix (Additional file 18: Dataset S9). All genes of both regions were included in the all-strain soft core with the 70% inclusion threshold. Thus, none of the genes were associated with pathotype and only the *paa* phenylacetic acid degradation pathway determinant was missing in the single-genome ECOR phylogroup B2 and E. This might have tipped the scales in the analysis of phylogroup A genomes by Goldstone and co-workers.

Several studies argue that MAEC strains from divergent phylogenetic backgrounds might use different VF subsets and virulence strategies to elicit bovine mastitis [10,30,34,35]. We tested this hypothesis exemplarily, by analyzing the 31 genes of the *fec*, *paa*, and *pga* regions for pathotype enrichment within the multi-genome phylogroups of our strain panel, A and B1. These three regions were detected as being essential in phylogroup A MAEC [35], but they have not been analyzed in other phylogroups. The 13 phylogroup A strains of our strain panel contain eleven MAEC and two commensal isolates. The ten strains of phylogenetic lineage B1 comprise four MAEC and six commensal strains. Although these are very low and uneven numbers regarding pathotype distribution, which might bias the results, this is still the first study to be able to perform such an analysis. No pathotype-enriched genes from the regions could be detected for the ECOR A genomes. All three regions were present in the group soft core (except for *paaB* in the unspecific category; Additional file 19: Dataset S10). In a similar way, the PGA biosynthesis and Fec regions were also mostly categorized into the group soft core of the analysis with B1 strains (Additional file 20: Dataset S11). Only *fecBCDE* were in the unspecific

category, because these genes are missing in the genomes of the commensal isolates RiKo 2331/09, O157:H43 T22, and W26. However, the whole seven-gene *pga* region was MAEC-enriched in our phylogroup B1 strain set, present in all four MAEC, but only in two of the six commensals. We want to stress that this result depends highly on the strain collection used and more bovine *E. coli* strains, especially commensals, from all available phylogroups need to be incorporated for an in-depth analysis. As all three regions are present in the all-strain soft core genome of our whole strain panel analysis, these results illustrate the drawbacks of inferring general observations from low numbers of strains (especially when focusing only on pathogenic strains) considering the genome plasticity of bovine *E. coli*.

Conclusions

This is the first publication to include a phylogenetically diverse bovine *E. coli* strain panel incorporating both MAEC and commensal isolates for genomic content comparisons. This study also includes the first two bovine *E. coli* genomes of finished quality [113] (MAEC 1303 and ECC-1470 [37]) and the largest collection of bovine *E. coli* commensals from fecal origin of udder-healthy cows [39]. This set of genomes enabled us to analyze differences in the gene content between MAEC and commensal strains in relation to the phylogenetic and genomic diversity of bovine *E. coli* in general. As we could not identify any genes associated with MAEC that were not also present in commensal strains or correlated with the strains' phylogenetic background, an MPEC pathotype characterized by specific VFs could not be defined. It is more likely that virulence-associated genes, which have been previously implicated in facilitating mastitis, have their principal function in colonization and persistence of the gastrointestinal habitat. Thus, like ExPEC, MAEC are facultative and opportunistic pathogens basically of naturally occurring commensal ("environmental") *E. coli* origin [14,21,24,25,29,31,54]. As a consequence, we propose to use the term mastitis-associated *E. coli* (MAEC) instead of mammary pathogenic *E. coli* (MPEC).

The genome content of certain bovine *E. coli* strains seems not to support the ability to elicit mastitis in udder-healthy cows as was shown in case of the commensal strain K71 [32]. The large pan-genome of bovine *E. coli* isolates offers many gene combinations to increase bacterial fitness by utilization of milk nutrients and evasion from the bovine innate immune system, thus resulting in sufficient bacterial intra-mammary growth and consequently infection of the mammary gland [10,94,105,114]. Isolates with an increased potential to cause mastitis can colonize the udder by chance depending on suitable environmental conditions and the cow's immune status. Our data also demonstrate, that there is no positive selection in MAEC for the presence of virulence-associated genes required for causing mastitis. This has implications for vaccine development and diagnostics. Reverse vaccinology may not be suitable for the identification of specific MAEC vaccine candidates, and the utilization of marker genes for improved diagnostics and prediction of the severity and outcome of an *E. coli* bovine mastitis might fail. Herd management and hygiene are still the two most important factors for preventing *E. coli* mastitis incidents. Several studies have shown a dramatic decrease in the bovine udder microbiome during mastitis, even after recovery [115–117]. It might be worthwhile to consider

alternative prevention strategies like strengthening the natural udder microbiota that competes with pathogens [118].

We urge the research community to not fall into the same trap with whole genome studies as with the previous typing studies. Mastitis researchers need to consolidate their efforts and, as Zadoks *et al.* eloquently put it, not to waste precious resources on “YATS” (yet another typing study) [5]. It is necessary to step away from the reductionist approach and adapt an integrated course of action by examining the host-pathogen interaction simultaneously. Modern techniques, like dual RNA-Seq of host and bacteria [119,120], Tn-Seq to test virulence association of genes *in vivo*, SNP analysis, proteomics, and metabolomics, are readily available to overcome the knowledge boundary in this field.

Methods

Bacterial strains, isolation, and published reference genome acquisition

All fourteen isolates in this study were collected using routine clinical practices from the bovine habitat [2,4]. Commensal strains were isolated from fecal samples of udder-healthy and mastitis-associated strains from the serous udder exudate of mastitis-afflicted cows. Mastitis strains were acquired from different veterinary diagnostic laboratories in the indicated countries, listed in the genomes feature overview table (Additional file 2: Table S2). Total DNA from overnight cultures for all strains was isolated with the MasterPure Complete DNA and RNA Purification Kit (Epicentre, Madison, WI, USA) according to the manufacturer's instructions. Additionally, eleven draft bovine-associated *E. coli* reference genomes were downloaded from NCBI to be used in the analyses. See Table 1 for the respective reference publications. The corresponding accession numbers are given in Additional file 2: Table S2.

Library preparation and sequencing

The strains with closed genomes, 1303 and ECC-1470, were sequenced as described in [37]. In short, both genomes were first sequenced with the 454 Titanium FLX genome sequencer with GS20 chemistry (Roche Life Science, Mannheim, Germany) in a whole-genome shotgun approach to 27.8-fold and overall 13.4-fold coverage, respectively (384,786 reads and 143,474,880 bases for *E. coli* 1303, 129,126 reads and 39,329,989 bases for *E. coli* ECC-1470). Strain ECC-1470 was also sequenced with a 6-kb insert paired-end (PE) 454 library (155,130 reads and 26,495,179 bases).

These two strains were additionally and the draft strains [39] solely sequenced with a 101-bp PE sequencing run on a HiScan SQ sequencer (Illumina, San Diego, CA, USA). For this purpose, sequencing libraries were prepared with Nextera XT chemistry. All Illumina raw reads were quality controlled with FastQC before and after trimming (v0.11.2; http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Median insert sizes of the PE Illumina

libraries were calculated with Picard's CollectInsertSizeMetrics (v1.124; http://broadinstitute.github.io/picard) after the raw reads were mapped onto the assembled contigs with Bowtie2 (v2.0.6) [121] (see used options below) and the mappings sorted with SAMtools (v0.1.19) [122]. Low quality 3' end of reads and Illumina adapter contaminations (--stringency 2) were trimmed with cutadapt (v1.6) with a Q20 Phred score cutoff and a minimum read length of 20 bp [123].

Assembly of the genomes

Both 454 read sets of the closed strains were *de novo* assembled with Newbler (Roche) (v2.0.00.20 for 1303 and v2.3 for ECC-1470) [124]. Additionally, these reads were assembled in a hybrid *de novo* approach in combination with the respective Illumina reads using MIRA (v3.4.0.1) [125]. MIRA assembly of the corresponding reads with a 26x fold 454 and 70x fold Illumina coverage resulted in the following statistics for 1303: 98 contigs >= 500 bp and an N50 of 165,271 bp. ECC-1470 was initially assembled with reads of a 12x fold 454 and 75x fold Illumina coverage: 88 contigs >= 500 bp and an N50 of 194,065 bp. Afterwards, each 454 Newbler assembly was combined with the respective hybrid assembly in Gap4 (v4.11.2) of the Staden software package [126]. The remaining gaps in the assembly were closed by primer walking via directed PCR and Sanger sequencing utilizing BigDye Terminator chemistry with ABI 3730 capillary sequencers. The sequences were processed with Pregap4 and loaded into the Gap4 databases. The closed genomes were edited to the "finished" standard [113].

The Illumina reads from the draft strains were each randomly subsampled to an approximate 70-fold coverage with seqtk (v1.0-r32; https://github.com/lh3/seqtk). Afterwards, the PE reads were *de novo* assembled with SPAdes (v3.1.1) with an iterative k-mer range of '-k 21,33,55,77' and option '--careful' to reduce the number of mismatches and insertion/deletions [127]. The following three steps were executed to check the assembled contigs: First, the reads used for the assemblies were mapped with Bowtie2 and its '--end-to-end', '--very-fast', and minimum (option '-I 0') and maximum ('-X 1000') PE insert size options. The resulting SAM files were then sorted by coordinates and converted to BAM files with SAMtools to calculate mapping statistics with QualiMap (v2.0) [128]. Only contigs >= 500 bp were retained, because smaller contigs often contain misassembled repeat sequences that cannot be resolved by the assembler. At last, the assembled contigs were ordered against the respective *E. coli* 1303 or *E. coli* ECC-1470 reference genomes, according to the ECOR phylogroup affiliation of the draft genomes. Contig ordering was done with ABACAS (v1.3.2) [129] running NUCmer (v3.1) and with order_fastx (v0.1) [130]. Assembly statistics were determined with QUAST (v3.2) [131] using NUCmer from the MUMmer package (v3.23) [132] for the 12 draft strains in this study and also for the 11 bovine-associated reference draft strains (with contigs >= 500 bp). All Sequence Read Archive (SRA) study accession numbers for the Illumina and 454 raw reads of the *E. coli* genomes of this study can be found in Additional file 1: Table S1. This file also includes the assembly statistics for all 23 bovine-associated *E. coli* draft genomes. The draft genomes of this study are in the "high-quality draft" standard [113].

All genomes of this study were scanned with BLASTN+ (v2.2.28) [133] for contamination with the Illumina phage PhiX spike-in control. Enterobacteria phage phiX174 genome (accession number: NC_001422.1) was used as query in the BLASTN+ runs.

Annotation of the genomes

All strains of this study were initially automatically annotated with Prokka (v1.9) [134] and the annotations subsequently supplemented with further databases. tRNAs were predicted with tRNAscan-SE (v1.3.1) [135].

For the *E. coli* 1303 and ECC-1470 chromosomes *E. coli* K-12 MG1655 (accession number: NC_000913.3) and for their F plasmids (p1303_109 and pECC-1470_100) *E. coli* K-12 CR63 F plasmid (NC_002483.1) were used as references in Prokka (option '--proteins'). 1303 P1 phage plasmid (p1303_95) was annotated with enterobacteria phage P1 (NC_005856.1) as reference. These initial annotations were manually curated with the Swiss-Prot, TrEMBL [136], IMG/ER [137], and Ecocyc databases [138]. Also, the Prodigal (v2.60) [139] ORF finding in Prokka was verified with a YACOP (v1) [140] ORF finding. Subsequently, the two annotations were compared to the highly curated reference annotation of strain MG1655 using the Artemis Comparison Tool (ACT) (v12.1.1) [141] with BLASTN+. With these comparisons manual curation was carried out with the tools Artemis (v15.1.1) [142] and tbl2tab (v0.1) [130]. Lastly, the annotations of *E. coli* strains 1303 and ECC-1470 were compared (ACT) and adapted to each other for a uniform annotation.

The high quality annotation of the *E. coli* 1303 genome was then used as reference for the ECOR phylogroup A strains and the ECC-1470 genome annotation for the ECOR B1 strains during the Prokka annotation of the 12 draft strains of this study. These annotations were further manually curated via ortholog/genome synteny analyses with the respective replicons of *E. coli* strains 1303 and ECC-1470 as references with Proteinortho (v5.11) [143,144] (see options below) and po2anno (v0.2) [130], ACT (v13.0.0) [141] with BLASTN+, and cat_seq (v0.1) [130]. At last, releases 1 (R1) and 2 (R2) of the Virulence Factors Database (VFDB) [145,146], and the ResFinder (v2.1) [147], VirulenceFinder (v1.2) [148], and SerotypeFinder (v1.0) [149] databases were used to refine the annotations with Artemis (v16.0.0) and tbl2tab (v0.2).

All eleven reference strains were also automatically reannotated with Prokka to have a uniform ORF-finding with Prodigal and facilitate comparative genomics. The draft genomes of D6-113.11 and D6-117_07.11 contain one contig each smaller than 200 bp. These two contigs were skipped by Prokka with the used option '--compliant'. The annotations of the references were shortly manually curated in the three putative virulence regions ETT2, Flag-2, and strain ECC-1470's T6SS/1 by comparisons to the 1303 and ECC-1470 genomes as mentioned above. GENBANK files for these reannotations were created with NCBI's tbl2asn (v24.3; https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/) with option '-V b' and can be found in Additional file 21: Dataset S12 and Additional file 22: Dataset S13.

For an overview of the annotations see the genome feature table created with genomes_feature_table (v0.5) [130] (Additional file 2: Table S2). This table also includes the reference *E. coli* genomes for the phylogenetic analysis (see below), however their annotation features are listed as downloaded from NCBI.

Phylogenetic analysis

For the phylogenetic analysis 39 additional reference *E. coli* strains (plus four *Shigella* spp. and one *Escherichia fergusonii* strain) were downloaded from NCBI with a wide variety of known pathotype and ECOR phylogroup affiliations. For the accession numbers see Additional file 2: Table S2. A WGA was done with the default parameter settings of Mugsy (v1.2.3) [150] and the combined 68 *E. coli* genomes (including plasmids) with *E. fergusonii* as outgroup. This resulted in an original alignment length of 3,764,795 bp. The MAF alignment file was further processed to contain only locally colinear blocks without gaps present in all aligned genomes utilizing the software suite Phylomark (v1.3) [151]. Phylomark in turn makes use of modules from Biopython (v1.63) [152] and bx-python (v0.7.1; https://github.com/bxlab/bx-python), and as a final step runs mothur (v1.22.2) [153]. After this treatment the resulting alignment length was 2,272,130 bp.

The concatenated and filtered alignment was then subjected to RAxML (v8.1.22) [154] to infer the best scoring ML phylogeny. RAxML was run with the GTRGAMMA generalized time-reversible (GTR) model of nucleotide evolution and GAMMA model of rate heterogeneity. 1,000 bootstrap resamplings were calculated with RAxML's rapid bootstrapping algorithm (option '-f a') for local support values. The resulting tree was visualized with Dendroscope (v3.4.4) [155]. This phylogeny was used to classify the bovine-associated strains into ECOR phylogroups according to the included reference strains (with a known phylogeny) and monophyletic clades.

The same procedure was followed including only the 25 bovine-associated *E. coli* strains. This resulted in a Mugsy alignment length of 4,312,845 bp and a filtered alignment length of 3,393,864 bp for RAxML. This tree was visualized with FigTree (v1.4.1; http://tree.bio.ed.ac.uk/software/figtree/) midpoint rooted.

STs were assigned with ecoli_mlst (v0.3) [130] according to the Achtman *E. coli* MLST scheme [13] employing NUCmer with default parameters. Ambiguous allele numbers for strains ECA-O157, ECA-727, and O157:H7 EDL933 were resolved with BLASTN+ by choosing the sequence allele with the highest identity in the MLST database. PHYLOViZ (v1.1) [156] was used to create a MST with the goeBURST algorithm [157] to classify the STs into CCs. CC numbers were allocated according to the Achtman *E. coli* MLST database. A CC is defined by STs that differ at maximal one locus/allele and are numbered by the founder of the CC, which is the ST with the highest number of neighboring single locus variants (SLVs). All allele, ST, and CC numbers can be found in Additional file 3: Table S3.

Detection of genomic islands and prophages, and generation of circular genome diagrams

GIs and prophages were predicted in the two closed genomes. GIs were predicted with the three prediction methods of IslandViewer 3 [158]: the two sequence composition methods SIGI-HMM [159] and IslandPath-DIMOB, and the comparative genomic prediction method IslandPick [160]. Only predicted GIs with a size greater than 8 kb were retained. Prophages were predicted with the PHAge Search Tool (PHAST) [161]. PHAST also evaluates the completeness and potential viability of prophage regions by classifying them as “intact”, “questionable”, or “incomplete”.

The GI and prophage predictions and their locations were evaluated manually by looking for mobility-associated genes, like integrases and transposons, toxin-antitoxin genes, restriction modification systems, and associated tRNAs using Artemis. The location, gene name (if available), locus tag, orientation, and product annotation was extracted for all genes included in the GI and prophage regions with Artemis.

Circular genome views were created with the BLAST Ring Image Generator (BRIG, v0.95) [162] using BLASTP+ (v2.2.28) [133] with a disabled low complexity filter (option ‘-seg no’) and upper/lower identity thresholds set to 90% and 70%, respectively. The location of the predicted GIs and prophages are visualized in these diagrams.

Identifying serotypes

The SerotypeFinder (v1.0) database from the Center for Genomic Epidemiology was used to determine serotypes *in silico* [149]. For some strains SerotypeFinder could not resolve the O- or H-antigen uniquely, in these cases both are listed.

Ortholog/paralog analysis

Orthologous and paralogous proteins in all 25 bovine-associated genomes were identified with Proteinortho (v5.11) [143,144] with a 1 x 10^{-5} E-value and 70% coverage/identity cutoffs. Proteinortho employs a bidirectional all-vs-all BLASTP+ (v2.2.29) approach using all predicted non-pseudo coding sequences, which were extracted from the genomes with cds_extractor (v0.7.1) and its option ‘-p’ [130]. Additionally, Proteinortho’s ‘-synteny’ option was used to activate the PoFF module enabling the utilization of genome synteny for improving ortholog detection. GFF3 files for this purpose were created with bp_genbank2gff3.pl from the BioPerl script collection (v1.6.924; https://github.com/bioperl/bioperl-live/tree/master/scripts/Bio-DB-GFF) [163]. Other non-default Proteinortho options used were a final local optimal Smith-Waterman alignment for BLASTP+ (‘-blastParameters=‘-use_sw_tback’’) recommended by Moreno-Hagelsieb and Latimer [164] and Ward and Moreno-Hagelsieb [165], ‘-selfblast’ for paralog detection, and ‘-singles’ to also report singletons. This resulted in a total number of 13,481 OGs from the overall 116,535 CDSs in the bovine-associated strain panel.

po2group_stats (v0.1.1) [130] was used to test for pathotype- (mastitis/commensal) or phylogroup-enriched (ECOR phylogroups A/B1/B2/E) OGs. The script can handle "fuzzy" presence/absence of OGs in genome groups (respective genome groups given with option '-g') by setting inclusion and exclusion cutoffs. 70% inclusion and 30% exclusion cutoffs were used for po2group_stats (options '-cut_i' and '-cut_e') with the genomes classified either according to pathotype (9 commensals and 16 MAEC) or phylogroup (13 ECOR A, 10 B1, and one for each B2 and E). OGs are pathotype- or phylogroup-enriched if they are minimally present in the genomes of one genome group (inclusion cutoff) and maximally in the genomes of all other groups (the other pathotype or phylogroups; exclusion cutoff). The 70%/30% inclusion/exclusion cutoffs amount to rounded 6/3 inclusion/exclusion genome cutoffs for the commensal isolates and 11/5 for the mastitis isolates. Similarly, the cutoffs in the phylogroups translate to rounded 9/4 genome inclusion/exclusion cutoffs for phylogroup A, 7/3 for B1, and 1/0 for the single genome groups B2 and E. According to these pathotype or phylogroup cutoffs, OGs are classified in "pathotype-/phylogroup-enriched", "-absent", "group soft core genome", "underrepresented", and "unspecific" (option '-u') categories. OGs that are present >= the inclusion cutoff in the genomes of all groups are categorized in the "group soft core genome" category. The "underrepresented" category includes OGs present in <= genomes than the exclusion cutoff in all groups. Finally, OGs that are present in more genomes than the exclusion but less than the inclusion cutoff in any group are categorized as "unspecific". For each OG po2group_stats extracts the locus tag and annotation of one representative protein from one *E. coli* strain panel genome of the group (or in the case of paralogs several representative proteins).

The resulting pathotype-/phylogroup-enriched and group soft core OG numbers were visualized in venn diagrams (po2group_stats option '-p') with the venn function of R package gplots (v3.0.1) [166]. Additionally, singletons (option '-s') were identified with po2group_stats.

In addition to the pathotype and phylogroup group soft core genomes calculated by po2group_stats, an "all-strain soft core genome" over all genomes with the 70% inclusion cutoff (rounded 18 genomes of the total 25) was determined. The all-strain soft core genome always includes more OGs than the pathotype/phylogroup group soft cores, because of the different number of groups the 70% inclusion cutoff is applied to. The difference originates from the inclusion of all OGs which are present in at least 70% of all genomes of each group in comparison to 70% of all genomes.

The resulting pathotype-enriched OGs were further evaluated by comparing their representative proteins to the representative proteins in the phylogroup-enriched categories and the all-strain soft core. The representative protein sequences were extracted from the respective GENBANK files with the locus tags included in the po2group_stats result files using cds_extractor (options '-p' and '-l'). Subsequently, the prot_finder pipeline with BLASTP+ was used, as described below in the VF workflow, with the pathotype-enriched representative proteins as queries (option '-q') and the phylogroup-enriched or all-strain/phylogroup soft core proteins as subjects (option '-s').

Finally, a gene content tree was calculated with the Proteinortho presence/absence matrix of OGs. First, the matrix was converted to a binary matrix, transposed with transpose_matrix (v0.1) [130], and then converted to FASTA format. This file was used to cluster the results by searching for the best scoring ML tree with RAxML's (v8.0.26) BINGAMMA module (binary substitution model with GAMMA model of rate heterogeneity) and 1000 resamplings. The clustering tree was visualized midpoint rooted with Figtree.

Screening of the genomes for known virulence factors

VF reference protein sequences for the used VF panel were collected by searching through all three releases of the VFDB (R1 core dataset with experimentally validated VFs [145], R2 comparative genomics dataset with intra-genera comparisons [146], and R3 VF centric dataset with inter-genera comparisons [167]) and reviewing the primary literature. Each VF was classified in one of twelve classes, like adhesion and invasion, chaperone-usher fimbriae, iron uptake systems, toxins, T6SS etc. For an overview of the VF panel see Additional file 12: Table S6. A focus was put on putative ExPEC VFs, because, to the authors knowledge, there is no adequate collection of these to be found and MAEC are considered to be ExPEC [3]. The protein sequences of the VFs, as well as detailed information how the VF panel was collected, and the respective reference publications can be found in the GitHub repository https://github.com/aleimba/ecoli_VF_collection (v0.1) [100].

The VF panel was used to assess the presence/absence of the 1,069 virulence-associated genes in the annotated bovine-associated strains with the prot_finder pipeline (v0.7.1) [130] using BLASTP+ (v2.2.29). The following non-default options were used for the prot_finder pipeline: 1 x 10^{-10} E-value cutoff ('-evalue 1e-10'), 70% query identity and coverage cutoffs (options '-i' and '-cov_q'), and the best BLASTP hits option ('-b'). This option includes only the hit with the highest identity for each subject CDS protein. A binary presence/absence matrix from these results was created with prot_binary_matrix (v0.6) and transpose_matrix (v0.1) [130]. As with the gene content tree, a ML RAxML BINGAMMA search was done to cluster the results in the binary matrix with 1,000 resamplings. The clustering tree was visualized midpoint rooted with Figtree and converted to an ultrameric cladogram. Additionally, the binary VF hit matrix was visualized with function heatmap.2 of the R package gplots and R package RColorBrewer (v1.1-2) [168]. The aforementioned cladogram was attached to this heatmap with R package ape (v3.4) [169]. The binary matrix, the cladogram NEWICK file, and the R script are included in Additional file 14: Dataset S6. The two resulting heatmaps were merged and edited in Inkscape.

The Perl script binary_groups_stats (v0.1) [130] (in the same manner as po2group_stats above) was used to detect VF genes enriched in the pathotypes or phylogroups of the binary matrix. Again inclusion and exclusion cutoffs were set to 70% and 30%, respectively. Venn diagrams visualized the number of pathotype- and phylogroup-enriched VF genes, as well as the group soft core VF sets. Also, an all-strain soft core VF set was calculated over the virulence-associated gene hits of all genomes with a 70% (18 genome) inclusion cutoff. Pathotype-

enriched VF proteins were compared to phylogroup-enriched or all-strain/phylogroup soft core VF proteins for evaluation.

The same prot_finder pipeline and binary_groups_stats workflow was also used for two putative MAEC-specific regions in ECOR phylogroup A genomes [35], which are not included in the VF panel. The first region is the biofilm-associated polysaccharide synthesis locus (*pgaABCD-ycdT-ymdE-ycdU*). The protein sequences from these genes were extracted from strain 1303 with cds_extractor (option '-l'). The locus tags are EC1303_c10400 to EC1303_c10440, EC1303_c10470, and EC1303_c10480. The second region encodes proteins involved in the phenylacetic acid degradation pathway (*feaRB-tynA-paaZABCDEFGHIJKXY*; MG1655 locus tags b1384 to b1400). The third region mentioned in the publication (the Fec uptake system, *fecIRABCDE*) is already included in the VF panel of this study. For this analysis the resulting binary BLASTP+ hit matrix (including the results for the *fec* genes) was also tested with binary_groups_stats for pathotype association within the ECOR A and B1 phylogroups of the bovine-associated strain panel. In detail, 11 mastitis and two commensal strains are in ECOR A (8/3 and 2/0 rounded genome inclusion/exclusion numbers for mastitis and commensal strains, respectively). ECOR B1 contains four mastitis and six commensal strains (3/1 and 4/2).

Analysis of large structural putative virulence regions

The composition of three putative large virulence regions ETT2, Flag-2, and strain ECC-1470 subtype i1 T6SS/1 was compared in more detail for the bovine-associated strain panel. To identify the corresponding contigs of the draft genomes the respective regions in *E. coli* strains 1303 and ECC-1470 were compared with ACT and BLASTN+ to the draft genomes. The identified draft contigs were optionally reversed with revcom_seq (v0.2), concatenated with cat_seq, and truncated with trunc_seq (v0.2) [130] to include two flanking core genome genes. ORFs that spanned contig borders in the concatenated sequence files were manually elongated or added with Artemis, these genes are marked by asterisks '*' in the figures. The genome comparison diagrams were created with Easyfig (v2.2.2) [170] using BLASTN+ with a maximal E-value of 0.001 and the genomes ordered according to the WGA phylogeny.

The same workflow was done for the antimicrobial multidrug resistance element of 1303 (AMR-SSuT in GI4) in comparison to the *E. coli* SSuT-25 AMR-SSuT element [78] (accession number: EF646764), the *E. coli* O157:H7 EC20020119 AMR-SSuT region (accession number: HQ018801) [79], and transposon Tn*10* of *Shigella flexneri* 2b plasmid R100 (accession number: AP000342).

General data generation and figure editing

Dendroscope was used to create tanglegrams between the cladograms of the bovine-associated strain panel WGA phylogeny, gene content, or VF clustering trees. All figures, created either with R (v3.2.5) [171] for the heatmap or venn diagrams, Dendroscope or FigTree for phylogenetic trees, PHYLOViZ for the MLST MST, or Easyfig for the genome diagrams were saved in SVG or PDF format for color editing, labelling, and scaling with Inkscape (v0.91)

without changing data representation. The only exception are the BRIG circular genome diagrams which were edited with Gimp (v2.8.16).

Additional files

Additional file 1: Table S1. This file includes the SRA study accession numbers for the Illumina and 454 raw reads of the 14 *E. coli* genomes of this study. Additionally it lists the assembly statistics for all 23 bovine-associated *E. coli* draft genomes. (XSLX 8 KB)

Additional file 2: Table S2. Genome feature table for the 64 *E. coli*, four *Shigella* spp., and the one *Escherichia fergusonii* genomes plus accession numbers. (XSLX 18 KB)

Additional file 3: Table S3. MLST allele profiles, ST and CC numbers for the 64 *E. coli* and four *Shigella* spp. strains. (XSLX 10 KB)

Additional file 4: Figure S1. Minimum spanning tree (MST) of the MLST results. **Figure S2.** Phylograms and tanglegrams for the 25 bovine-associated *E. coli* genomes based on WGA core genome, gene and VF content. **Figure S3.** Circular genome diagrams for the MAEC 1303 and ECC-1470 replicons with GI and prophage positions. **Figure S4.** Heatmap of VF presence/absence, including gene names/locus tags. **Figure S5.** Gene organization of the AMR-SSuT/Tn*10* gene cluster. **Figure S6.** Gene organization of the Flag-2 and ECC-1470 T6SS/1 gene clusters. (PDF 16 MB)

Additional file 5: Table S4. Binary presence/absence matrix of 13,481 OGs in the 25 bovine-associated *E. coli* genomes. (XSLX 761 KB)

Additional file 6: Dataset S1. Singleton OGs in the 25 bovine-associated *E. coli* genomes. (XSLX 248 KB)

Additional file 7: Dataset S2. This file includes the pathotype-enriched OGs (MAEC or commensal isolates) with a 70% inclusion and 30% exclusion cutoff and their potential association with phylogroup-enriched categories or soft core genomes. It also specifies the pathotype group soft core genome and OGs classified as underrepresented and unspecific. For each OG the locus tag and annotation of one representative protein from one *E. coli* genome of the group is shown (or in the case of paralogs several representative proteins). (XSLX 520 KB)

Additional file 8: Table S5. All-strain soft core genome with 70% inclusion cutoff. (XSLX 185 KB)

Additional file 9: Dataset S3. Phylogroup-enriched OGs (A, B1, B2, or E), phylogroup group soft core, and underrepresented and unspecific OGs. (XSLX 532 KB)

Additional file 10: Dataset S4. Predicted GIs and prophages of MAEC 1303. (XSLX 68 KB)

Additional file 11: Dataset S5. Predicted GIs and prophages of MAEC ECC-1470. (XSLX 46 KB)

Additional file 12: Table S6. This file contains the overview of the VF panel. Presence ('1') and absence ('0') of the virulence-associated genes in the 25 bovine-associated *E. coli* genomes is indicated in column "present_in_strain_panel". Virulence-associated genes were collected from the Virulence Factors Database (VFDB) or from the primary literature ('own' in column "source"). (XSLX 55 KB)

Additional file 13: Table S7. BLASTP+ hit results for the VF panel in the 25 bovine-associated *E. coli* genomes. (XSLX 365 KB)

Additional file 14: Dataset S6. This zip archive contains the binary presence/absence matrix of virulence-associated genes in the 25 bovine-associated *E. coli* genomes, the VF content clustering cladogram in NEWICK format, and the R script to create the heatmaps in Figure 4 and S4. (ZIP 6 KB)

Additional file 15: Table S8. All-strain soft core VF set with 70% inclusion cutoff. (XSLX 10 KB)

Additional file 16: Dataset S7. This file includes the pathotype-enriched virulence-associated genes (MAEC or commensal isolates) with a 70% inclusion and 30% exclusion cutoff and their potential association with phylogroup-enriched categories or soft core genomes. It also specifies the pathotype group soft core VF set and virulence-associated genes classified as underrepresented and unspecific. (XSLX 25 KB)

Additional file 17: Dataset S8. Phylogroup-enriched virulence-associated genes (A, B1, B2, or E), phylogroup group soft core VF set, and underrepresented and unspecific virulence-associated genes. (XSLX 39 KB)

Additional file 18: Dataset S9. BLASTP+ hit results for the *pga* and *paa* gene regions and binary presence/absence matrix in the 25 bovine-associated *E. coli* genomes. (XSLX 37 KB)

Additional file 19: Dataset S10. Pathotype group soft core and unspecific categorisation of the *fec*, *paa*, and *pga* gene regions in the 13 phylogroup A bovine-associated *E. coli* genomes. (XSLX 9 KB)

Additional file 20: Dataset S11. Pathotype-enriched, group soft core, and unspecific categorisation of the *fec*, *paa*, and *pga* gene regions in the ten phylogroup B1 bovine-associated *E. coli* genomes. (XSLX 9 KB)

Additional file 21: Dataset S12. This zip archive contains the GENBANK files with the reannotations of five of the eleven reference bovine-associated *E. coli* genomes. Included are *E. coli* strains AA86, D6-113.11, D6-117.07, D6-117.29, and ECA-727. (ZIP 16 MB)

Additional file 22: Dataset S13. This zip archive contains the GENBANK files with the reannotations of six of the eleven reference bovine-associated *E. coli* genomes. Included are *E. coli* strains ECA-O157, ECC-Z, O32:H37 P4, P4-NR, O157:H43 T22, and W26. (ZIP 19 MB)

Abbreviations

AMR-SSuT: antimicrobial multidrug resistance to streptomycin, sulfonamide, and tetracycline; APEC: avian pathogenic *E. coli*; BRIG: BLAST Ring Image Generator; CC: clonal complex; CDS: coding DNA sequence; CU: chaperone usher pathway fimbriae; EAEC: enteroaggregative *E. coli*; ECOR: *E. coli* Reference; ECP: *E. coli* common pilus; EHEC: enterohaemorrhagic *E. coli*; EPEC: enteropathogenic *E. coli*; ETEC: enterotoxigenic *E. coli*; ETT2: *E. coli* type III secretion system 2; ExPEC: extraintestinal pathogenic *E. coli*; Fec: ferric iron(III)-dicitrate uptake system; FF: fitness factor; Flag-1: *E. coli* peritrichous flagella 1 gene cluster; Flag-2: *E. coli* lateral flagella 2 gene cluster; G4C: group 4 capsule; GI: genomic island; GTR: generalized time-reversible; HGT: horizontal gene transfer; IPEC: intestinal pathogenic *E. coli*; IS: insertion sequence; LEE: locus of enterocyte effacement; LPS: lipopolysaccharide; MAEC: mastitis-associated *E. coli*; MGE: mobile genetic element; ML: maximum likelihood; MNEC: newborn meningitis-associated *E. coli*; MPEC: mammary pathogenic *E. coli*; MST: minimum spanning tree; NET: neutrophil extracellular trap; OG: orthologous group; OMP: outer membrane protein; ORF: open reading frame; PAI: pathogenicity island; PAMP: pathogen-associated molecular pattern; PE: paired-end; PHAST: PHAge Search Tool; PMN: polymorphonuclear neutrophil; PTS: phosphotransferase system; Rhs: rearrangement hotspot; SLV: single locus variant; SPATE: serine protease autotransporters of *Enterobacteriaceae*; SRA: Sequence Read Archive; ST: sequence type; T2SS: type II secretion system; T3SS: type III secretion system; T5SS: type V secretion system; T6SS: type VI secretion system; TLR: Toll-like receptor; UPEC: uropathogenic *E. coli*; VFDB: Virulence Factors Database; VF: virulence factor; WGA: whole genome nucleotide alignment.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All raw reads of this study (454 and Illumina) can be accessed from NCBI's SRA. The corresponding SRA study accession numbers are listed in Additional File 1: Table S1. The assembled and annotated genomes of this study have been deposited at DDBJ/ENA/GenBank

under the accession numbers listed in Additional file 2: Table S2. The reannotated sequence files of the eleven bovine-associated reference *E. coli* are available in Additional file 21: Dataset S12 and Additional file 22: Dataset S13. The *E. coli* VF panel is available in the GitHub repository ecoli_VF_collection, https://github.com/aleimba/ecoli_VF_collection [100].

The R script used in this study can be found in Additional file 14: Dataset S6. R packages used are mentioned in the methods chapter. Perl scripts are stored in GitHub repository bac-genomics-scripts, https://github.com/aleimba/bac-genomics-scripts [130]. These are licensed under GNU GPLv3. Many of these depend on the BioPerl (v1.006923) module collection [163]. All other data sets supporting the results of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work, including the efforts of AL and UD, was funded by Deutsche Forschungsgemeinschaft (DFG) (DO 789/3-1 and DO 789/4-1).

Authors' contributions

Conceptualization: AL UD.
Data curation: AL AP.
Formal analysis: AL JV.
Funding acquisition: UD.
Investigation: AL.
Methodology: AL.
Project administration: AL.
Resources: AL RD UD.
Software: AL.
Supervision: AL UD.
Validation: AL.
Visualization: AL.
Writing – original draft: AL.
Writing – review & editing: AL UD JV RD AP.
All authors read and approved the final manuscript.

Acknowledgements

Thanks goes to Alan McNally for hosting Andreas Leimbach for a week and showing him some of the comparative genomics ropes. We would also like to thank David E. Kerr, Wolfram Petzl, Ynte Schukken, Nahum Shpigel, Olga Wellnitz, Lothar Wieler, and Holm Zerbe for supplying strains.

References

1. Hogeveen H, Huijps K, Lam TJ. Economic aspects of mastitis: new developments. N. Z. Vet. J. 2011;59:16–23.

2. Petzl W, Zerbe H, Günther J, Yang W, Seyfert H-M, Nürnberg G, et al. *Escherichia coli*, but not *Staphylococcus aureus* triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow. Vet. Res. 2008;39:18.

3. Shpigel NY, Elazar S, Rosenshine I. Mammary pathogenic *Escherichia coli*. Curr. Opin. Microbiol. 2008;11:60–5.

4. Dogan B, Klaessig S, Rishniw M, Almeida RA, Oliver SP, Simpson K, et al. Adherent and invasive *Escherichia coli* are associated with persistent bovine mastitis. Vet. Microbiol. 2006;116:270–82.

5. Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J. Mammary Gland Biol. Neoplasia. 2011;16:357–72.

6. Dogan B, Rishniw M, Bruant G, Harel J, Schukken YH, Simpson KW. Phylogroup and *lpfA* influence epithelial invasion by mastitis associated *Escherichia coli*. Vet. Microbiol. 2012;159:163–70.

7. Burvenich C, Van Merris V, Mehrzad J, Diez-Fraile A, Duchateau L. Severity of *E. coli* mastitis is mainly determined by cow factors. Vet. Res. 2003;34:521–64.

8. Porcherie A, Cunha P, Trotereau A, Roussel P, Gilbert FB, Rainard P, et al. Repertoire of *Escherichia coli* agonists sensed by innate immunity receptors of the bovine udder and mammary epithelial cells. Vet. Res. 2012;43:14.

9. Houser BA, Donaldson SC, Padte R, Sawant AA, DebRoy C, Jayarao BM. Assessment of phenotypic and genotypic diversity of *Escherichia coli* shed by healthy lactating dairy cattle. Foodborne Pathog. Dis. 2008;5:41–51.

10. Blum S, Heller ED, Krifucks O, Sela S, Hammer-Muntz O, Leitner G. Identification of a bovine mastitis *Escherichia coli* subset. Vet. Microbiol. 2008;132:135–48.

11. Blum SE, Leitner G. Genotyping and virulence factors assessment of bovine mastitis *Escherichia coli*. Vet. Microbiol. 2013;163:305–12.

12. Díaz E, Ferrández A, Prieto MA, García JL. Biodegradation of aromatic compounds by *Escherichia coli*. Microbiol. Mol. Biol. Rev. 2001;65:523–69.

13. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 2006;60:1136–51.

14. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. 2010;8:207–17.

15. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet.

2009;5:e1000344.

16. Chaudhuri RR, Henderson IR. The evolution of the *Escherichia coli* phylogeny. Infect. Genet. Evol. 2012;12:214–26.

17. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2004;2:123–40.

18. Johnson TJ, Wannemuehler Y, Johnson SJ, Stell AL, Doetkott C, Johnson JR, et al. Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens. Appl. Environ. Microbiol. 2008;74:7043–50.

19. Croxen MA, Finlay BB. Molecular mechanisms of *Escherichia coli* pathogenicity. Nat. Rev. Microbiol. 2010;8:26–38.

20. Köhler C-D, Dobrindt U. What defines extraintestinal pathogenic *Escherichia coli*? Int. J. Med. Microbiol. 2011;301:642–7.

21. Leimbach A, Hacker J, Dobrindt U. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. Curr. Top. Microbiol. Immunol. 2013;358:3–32.

22. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin. Microbiol. Rev. 2013;26:822–80.

23. Dobrindt U. (Patho-)Genomics of *Escherichia coli*. Int. J. Med. Microbiol. 2005;295:357–71.

24. Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, et al. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol. Biol. Evol. 2007;24:2373–84.

25. Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. J. Bacteriol. 2010;192:4885–93.

26. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. BMC Genomics. 2012;13:577.

27. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr. Opin. Genet. Dev. 2005;15:589–94.

28. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. Nat. Rev. Microbiol. 2004;2:414–24.

29. Wenz JR, Barrington GM, Garry FB, Ellis RP, Magnuson RJ. *Escherichia coli* isolates' serotypes, genotypes, and virulence genes and clinical coliform mastitis severity. J. Dairy Sci. 2006;89:3408–12.

30. Fernandes JBC, Zanardo LG, Galvão NN, Carvalho IA, Nero LA, Moreira MAS. *Escherichia coli* from clinical mastitis: serotypes and virulence factors. J. Vet. Diagn. Invest. 2011;23:1146–52.

31. Suojala L, Pohjanvirta T, Simojoki H, Myllyniemi A-L, Pitkälä A, Pelkonen S, et al. Phylogeny, virulence factors and antimicrobial susceptibility of *Escherichia coli* isolated in clinical bovine mastitis. Vet. Microbiol. 2011;147:383–8.

32. Blum SE, Heller ED, Sela S, Elad D, Edery N, Leitner G. Genomic and Phenomic Study of Mammary Pathogenic *Escherichia coli*. PLoS One. 2015;10:e0136387.

33. Richards VP, Lefébure T, Pavinski Bitar PD, Dogan B, Simpson KW, Schukken YH, et al. Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic *Escherichia coli*. PLoS One. 2015;10:e0119799.

34. Kempf F, Slugocki C, Blum SE, Leitner G, Germon P. Genomic Comparative Study of Bovine Mastitis *Escherichia coli*. PLoS One. 2016;11:e0147954.

35. Goldstone RJ, Harris S, Smith DGE. Genomic content typifying a prevalent clade of bovine mastitis-associated *Escherichia coli*. Sci. Rep. 2016;6:30115.

36. Yi H, Cho Y-J, Hur H-G, Chun J. Genome sequence of *Escherichia coli* AA86, isolated from cow feces. J. Bacteriol. 2011;193:3681.

37. Leimbach A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R, et al. Complete genome sequences of *Escherichia coli* strains 1303 and ECC-1470 isolated from bovine mastitis. Genome Announc. 2015;3:e00182–15.

38. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, et al. The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. BMC Genomics. 2011;12:9.

39. Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, et al. Whole-Genome Draft Sequences of Six Commensal Fecal and Six Mastitis-Associated *Escherichia coli* Strains of Bovine Origin. Genome Announc. 2016;4:e00753–16.

40. Kempf F, Loux V, Germon P. Genome sequences of two bovine mastitis-causing *Escherichia coli* strains. Genome Announc. 2015;3:e00259–15.

41. Tóth I, Schmidt H, Kardos G, Lancz Z, Creuzburg K, Damjanova I, et al. Virulence genes and molecular typing of different groups of *Escherichia coli* O157 strains in cattle. Appl. Environ. Microbiol. 2009;75:6282–91.

42. Sváb D, Horváth B, Szucs A, Maróti G, Tóth I. Draft Genome Sequence of an *Escherichia coli* O157:H43 Strain Isolated from Cattle. Genome Announc. 2013;1:e00263–13.

43. Sváb D, Bálint B, Maróti G, Tóth I. Cytolethal distending toxin producing *Escherichia coli* O157:H43 strain T22 represents a novel evolutionary lineage within the O157 serogroup. Infect. Genet. Evol. 2016;46:110–7.

44. Blum S, Sela N, Heller ED, Sela S, Leitner G. Genome analysis of bovine-mastitis-associated *Escherichia coli* O32:H37 strain P4. J. Bacteriol. 2012;194:3732.

45. Kim M, Yi H, Cho Y-J, Jang J, Hur H-G, Chun J. Draft genome sequence of *Escherichia coli* W26, an enteric strain isolated from cow feces. J. Bacteriol. 2012;194:5149–50.

46. Cooper KK, Mandrell RE, Louie JW, Korlach J, Clark TA, Parker CT, et al. Comparative

genomics of enterohemorrhagic *Escherichia coli* O145:H28 demonstrates a common evolutionary lineage with *Escherichia coli* O157:H7. BMC Genomics. 2014;15:17.

47. Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, et al. A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. Infect. Immun. 2011;79:950–60.

48. Clermont O, Olier M, Hoede C, Diancourt L, Brisse S, Keroudean M, et al. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. Infect. Genet. Evol. 2011;11:654–62.

49. Ghanbarpour R, Oswald E. Phylogenetic distribution of virulence genes in *Escherichia coli* isolated from bovine mastitis in Iran. Res. Vet. Sci. 2010;88:6–10.

50. Liu Y, Liu G, Liu W, Liu Y, Ali T, Chen W, et al. Phylogenetic group, virulence factors and antimicrobial resistance of *Escherichia coli* associated with bovine mastitis. Res. Microbiol. 2014;165:273–7.

51. Gordon DM, Clermont O, Tolley H, Denamur E. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ. Microbiol. 2008;10:2484–96.

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

53. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. 2013;5:58–65.

54. Tourret J, Denamur E. Population Phylogenomics of Extraintestinal Pathogenic *Escherichia coli*. Microbiol Spectr. 2016;4:UTI – 0010–2012.

55. Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics. 2012;13:256.

56. Leopold SR, Sawyer SA, Whittam TS, Tarr PI. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. BMC Evol. Biol. 2011;11:183.

57. McNally A, Cheng L, Harris SR, Corander J. The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. Genome Biol. Evol. 2013;5:699–710.

58. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 2008;190:6881–93.

59. Raetz CRH, Whitfield C. Lipopolysaccharide endotoxins. Annu. Rev. Biochem. 2002;71:635–700.

60. Garcia EC, Brumbaugh AR, Mobley HLT. Redundancy and specificity of *Escherichia coli* iron acquisition systems during urinary tract infection. Infect. Immun. 2011;79:1225–35.

61. Huja S, Oren Y, Trost E, Brzuszkiewicz E, Biran D, Blom J, et al. Genomic avenue to avian colisepticemia. MBio. 2015;6:e01681–14.

62. Poole SJ, Diner EJ, Aoki SK, Braaten BA, t'Kint de Roodenbeke C, Low DA, et al. Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems. PLoS Genet. 2011;7:e1002217.

63. Koskiniemi S, Lamoureux JG, Nikolakakis KC, t'Kint de Roodenbeke C, Kaplan MD, Low DA, et al. Rhs proteins from diverse bacteria mediate intercellular competition. Proc. Natl. Acad. Sci. U. S. A. 2013;110:7032–7.

64. Wang X, Preston JF, Romeo T. The *pgaABCD* locus of *Escherichia coli* promotes the synthesis of a polysaccharide adhesin required for biofilm formation. J. Bacteriol. 2004;186:2724–34.

65. Cerca N, Jefferson KK. Effect of growth conditions on poly-*N*-acetylglucosamine expression and biofilm formation in *Escherichia coli*. FEMS Microbiol. Lett. 2008;283:36–41.

66. Vuong C, Kocianova S, Voyich JM, Yao Y, Fischer ER, DeLeo FR, et al. A crucial role for exopolysaccharide modification in bacterial biofilm formation, immune evasion, and virulence. J. Biol. Chem. 2004;279:54881–6.

67. Kropec A, Maira-Litran T, Jefferson KK, Grout M, Cramton SE, Götz F, et al. Poly-*N*-acetylglucosamine production in *Staphylococcus aureus* is essential for virulence in murine models of systemic infection. Infect. Immun. 2005;73:6868–76.

68. Venketaraman V, Lin AK, Le A, Kachlany SC, Connell ND, Kaplan JB. Both leukotoxin and poly-*N*-acetylglucosamine surface polysaccharide protect *Aggregatibacter actinomycetemcomitans* cells from macrophage killing. Microb. Pathog. 2008;45:173–80.

69. Gomes F, Saavedra MJ, Henriques M. Bovine mastitis disease/pathogenicity: evidence of the potential role of microbial biofilms. Pathog. Dis. 2016;74:ftw006.

70. Hoffman JA, Badger JL, Zhang Y, Huang SH, Kim KS. *Escherichia coli* K1 *aslA* contributes to invasion of brain microvascular endothelial cells in vitro and in vivo. Infect. Immun. 2000;68:5062–7.

71. Ren C-P, Chaudhuri RR, Fivian A, Bailey CM, Antonio M, Barnes WM, et al. The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. J. Bacteriol. 2004;186:3547–60.

72. Yao Y, Xie Y, Perace D, Zhong Y, Lu J, Tao J, et al. The type III secretion system is involved in the invasion and intracellular survival of *Escherichia coli* K1 in human brain microvascular endothelial cells. FEMS Microbiol. Lett. 2009;300:18–24.

73. Cheng D, Zhu S, Su Z, Zuo W, Lu H. Prevalence and isoforms of the pathogenicity island ETT2 among *Escherichia coli* isolates from colibacillosis in pigs and mastitis in cows. Curr. Microbiol. 2012;64:43–9.

74. Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ. Bacterial genomics and pathogen evolution. Cell. 2006;124:703–14.

75. Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL. The association of virulence factors with genomic islands. PLoS One. 2009;4:e8094.

76. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 2011;13:36–46.

77. Łobocka MB, Rose DJ, Plunkett G, Rusin M, Samojedny A, Lehnherr H, et al. Genome of bacteriophage P1. J. Bacteriol. 2004;186:7032–68.

78. Khachatryan AR, Besser TE, Call DR. The streptomycin-sulfadiazine-tetracycline antimicrobial resistance element of calf-adapted *Escherichia coli* is widely distributed among isolates from Washington state cattle. Appl. Environ. Microbiol. 2008;74:391–5.

79. Ziebell K, Johnson RP, Kropinski AM, Reid-Smith R, Ahmed R, Gannon VP, et al. Gene cluster conferring streptomycin, sulfonamide, and tetracycline resistance in *Escherichia coli* O157:H7 phage types 23, 45, and 67. Appl. Environ. Microbiol. 2011;77:1900–3.

80. Feng L, Liu B, Liu Y, Ratiner YA, Hu B, Li D, et al. A genomic islet mediates flagellar phase variation in *Escherichia coli* strains carrying the flagellin-specifying locus *flk*. J. Bacteriol. 2008;190:4470–7.

81. Ratiner YA, Sihvonen LM, Liu Y, Wang L, Siitonen A. Alteration of flagellar phenotype of *Escherichia coli* strain P12b, the standard type strain for flagellar antigen H17, possessing a new non-*fliC* flagellin gene *flnA*, and possible loss of original flagellar phenotype and genotype in the course of subculturing through semisolid media. Arch. Microbiol. 2010;192:267–78.

82. Liu B, Hu B, Zhou Z, Guo D, Guo X, Ding P, et al. A novel non-homologous recombination-mediated mechanism for *Escherichia coli* unilateral flagellar phase variation. Nucleic Acids Res. 2012;40:4530–8.

83. Tominaga A. Characterization of six flagellin genes in the H3, H53 and H54 standard strains of *Escherichia coli*. Genes Genet. Syst. 2004;79:1–8.

84. Lügering A, Benz I, Knochenhauer S, Ruffing M, Schmidt MA. The Pix pilus adhesin of the uropathogenic *Escherichia coli* strain X2194 (O2 : K(-): H6) is related to Pap pili but exhibits a truncated regulatory region. Microbiology. 2003;149:1387–97.

85. Schneider G, Dobrindt U, Brüggemann H, Nagy G, Janke B, Blum-Oehler G, et al. The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536. Infect. Immun. 2004;72:5993–6001.

86. Hejnova J, Pages D, Rusniok C, Glaser P, Sebo P, Buchrieser C. Specific regions of genome plasticity and genetic diversity of the commensal *Escherichia coli* A0 34/86. Int. J. Med. Microbiol. 2006;296:541–6.

87. Barondess JJ, Beckwith J. *bor* gene of phage lambda, involved in serum resistance, encodes a widely conserved outer membrane lipoprotein. J. Bacteriol. 1995;177:1247–53.

88. Johnson TJ, Wannemuehler YM, Nolan LK. Evolution of the *iss* gene in *Escherichia coli*. Appl. Environ. Microbiol. 2008;74:2360–9.

89. Kukkonen M, Korhonen TK. The omptin family of enterobacterial surface

proteases/adhesins: from housekeeping in *Escherichia coli* to systemic spread of *Yersinia pestis*. Int. J. Med. Microbiol. 2004;294:7–14.

90. Hwang B-Y, Varadarajan N, Li H, Rodriguez S, Iverson BL, Georgiou G. Substrate specificity of the *Escherichia coli* outer membrane protease OmpP. J. Bacteriol. 2007;189:522–30.

91. Haiko J, Suomalainen M, Ojala T, Lähteenmäki K, Korhonen TK. Invited review: Breaking barriers--attack on innate immune defences by omptin surface proteases of enterobacterial pathogens. Innate Immun. 2009;15:67–80.

92. Bradley A. Bovine mastitis: an evolving disease. Vet. J. 2002;164:116–28.

93. Kornalijnslijper JE, van Werven T, van den Broek J, Daemen AJ, Niewold TA, Rutten VP, et al. In vitro growth of mastitis-inducing *Escherichia coli* in milk and milk fractions of dairy cows. Vet. Microbiol. 2003;91:125–34.

94. Kornalijnslijper JE, van Werven T, Daemen AJ, Niewold TA, Rutten VP, Noordhuizen-Stassen EN. Bacterial growth during the early phase of infection determines the severity of experimental *Escherichia coli* mastitis in dairy cows. Vet. Microbiol. 2004;101:177–86.

95. Boulanger V, Bouchard L, Zhao X, Lacasse P. Induction of nitric oxide production by bovine mammary epithelial cells and blood leukocytes. J. Dairy Sci. 2001;84:1430–7.

96. Kaipainen T, Pohjanvirta T, Shpigel NY, Shwimmer A, Pyörälä S, Pelkonen S. Virulence factors of *Escherichia coli* isolated from bovine clinical mastitis. Vet. Microbiol. 2002;85:37–46.

97. Döpfer D, Almeida RA, Lam TJ, Nederbragt H, Oliver SP, Gaastra W. Adhesion and invasion of *Escherichia coli* from single and recurrent clinical cases of bovine mastitis in vitro. Vet. Microbiol. 2000;74:331–43.

98. Schukken YH, Günther J, Fitzpatrick J, Fontaine MC, Goetze L, Holst O, et al. Host-response patterns of intramammary infections in dairy cows. Vet. Immunol. Immunopathol. 2011;144:270–89.

99. Günther J, Koy M, Berthold A, Schuberth H-J, Seyfert H-M. Comparison of the pathogen species-specific immune response in udder derived cell types and their models. Vet. Res. 2016;47:22.

100. Leimbach A. ecoli_VF_collection: v0.1. Zenodo. 2016. http://dx.doi.org/10.5281/zenodo.56686

101. Lippolis JD, Brunelle BW, Reinhardt TA, Sacco RE, Thacker TC, Looft TP, et al. Differential Gene Expression of Three Mastitis-Causing *Escherichia coli* Strains Grown under Planktonic, Swimming, and Swarming Culture Conditions. mSystems. 2016;1:e00064–16.

102. Mike LA, Smith SN, Sumner CA, Eaton KA, Mobley HLT. Siderophore vaccine conjugates protect against uropathogenic *Escherichia coli* urinary tract infection. Proc. Natl. Acad. Sci. U. S. A. 2016;113:13468–73.

103. Xie Y, Kim KJ, Kim KS. Current concepts on *Escherichia coli* K1 translocation of the blood-brain barrier. FEMS Immunol. Med. Microbiol. 2004;42:271–9.

104. Nemeth J, Muckle CA, Lo RY. Serum resistance and the *traT* gene in bovine mastitis-

causing *Escherichia coli*. Vet. Microbiol. 1991;28:343–51.

105. Rainard P, Riollet C. Innate immunity of the bovine mammary gland. Vet. Res. 2006;37:369–400.

106. Ren C-P, Beatson SA, Parkhill J, Pallen MJ. The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. J. Bacteriol. 2005;187:1430–40.

107. Li J, Yao Y, Xu HH, Hao L, Deng Z, Rajakumar K, et al. SecReT6: a web-based resource for type VI secretion systems found in bacteria. Environ. Microbiol. 2015;17:2196–202.

108. Journet L, Cascales E. The Type VI Secretion System in *Escherichia coli* and Related Species. EcoSal Plus. 2016;7:ESP – 0009–2015.

109. Lehtolainen T, Pohjanvirta T, Pyörälä S, Pelkonen S. Association between virulence factors and clinical course of *Escherichia coli* mastitis. Acta Vet. Scand. 2003;44:203–5.

110. Fairbrother J-H, Dufour S, Fairbrother JM, Francoz D, Nadeau É, Messier S. Characterization of persistent and transient *Escherichia coli* isolates recovered from clinical mastitis episodes in dairy cows. Vet. Microbiol. 2015;176:126–33.

111. Johnson TJ, Johnson SJ, Nolan LK. Complete DNA sequence of a ColBM plasmid from avian pathogenic *Escherichia coli* suggests that it evolved from closely related ColV virulence plasmids. J. Bacteriol. 2006;188:5975–83.

112. Johnson TJ, Siek KE, Johnson SJ, Nolan LK. DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains. J. Bacteriol. 2006;188:745–58.

113. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, et al. Genome project standards in a new era of sequencing. Science. 2009;326:236–7.

114. Vangroenweghe F, Rainard P, Paape M, Duchateau L, Burvenich C. Increase of *Escherichia coli* inoculum doses induces faster innate immune response in primiparous cows. J. Dairy Sci. 2004;87:4132–44.

115. Oikonomou G, Machado VS, Santisteban C, Schukken YH, Bicalho RC. Microbial diversity of bovine mastitic milk as described by pyrosequencing of metagenomic 16s rDNA. PLoS One. 2012;7:e47671.

116. Falentin H, Rault L, Nicolas A, Bouchard DS, Lassalas J, Lamberton P, et al. Bovine teat microbiome analysis revealed reduced alpha diversity and significant changes in taxonomic profiles in quarters with a history of mastitis. Front. Microbiol. 2016;7:480.

117. Ganda EK, Bisinotto RS, Lima SF, Kronauer K, Decter DH, Oikonomou G, et al. Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin. Sci. Rep. 2016;6:37565.

118. Bouchard DS, Seridan B, Saraoui T, Rault L, Germon P, Gonzalez-Moreno C, et al. Lactic Acid Bacteria Isolated from Bovine Mammary Microbiota: Potential Allies against Bovine Mastitis. PLoS One. 2015;10:e0144831.

119. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nat. Rev. Microbiol. 2012;10:618–30.

120. Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, Schulte LN, et al. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. Nature. 2016;529:496–501.

121. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012;9:357–9.

122. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

123. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10.

124. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.

125. Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. German conference on bioinformatics. bioinfo.de. 1999. http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html

126. Staden R, Beal KF, Bonfield JK. The Staden package, 1998. Methods Mol. Biol. 2000;132:115–30.

127. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 2012;19:455–77.

128. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4.

129. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25:1968–9.

130. Leimbach A. bac-genomics-scripts: Bovine *E. coli* mastitis comparative genomics edition. Zenodo. 2016. http://dx.doi.org/10.5281/zenodo.215824

131. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.

132. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

133. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

134. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

135. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955–64.

136. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42:D191–8.

137. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics. 2009;25:2271–8.

138. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, et al. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 2013;41:D605–12.

139. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

140. Tech M, Merkl R. YACOP: Enhanced gene prediction obtained by a combination of existing methods. In Silico Biol. 2003;3:441–51.

141. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 2005;21:3422–3.

142. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16:944–5.

143. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011;12:124.

144. Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, et al. Orthology detection combining clustering and synteny for very large datasets. PLoS One. 2014;9:e105015.

145. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33:D325–8.

146. Yang J, Chen L, Sun L, Yu J, Jin Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res. 2008;36:D539–42.

147. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. 2012;67:2640–4.

148. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J. Clin. Microbiol. 2014;52:1501–10.

149. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy *in silico* serotyping of *Escherichia coli* using whole genome sequencing (WGS) data. J. Clin. Microbiol. 2015;53:2410–26.

150. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27:334–42.

151. Sahl JW, Matalka MN, Rasko DA. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl. Environ. Microbiol. 2012;78:4884–92.

152. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.

153. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 2009;75:7537–41.

154. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

155. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 2012;61:1061–7.

156. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics. 2012;13:87.

157. Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009;10:152.

158. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 2015;43:W104–8.

159. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics. 2006;7:142.

160. Langille MG, Hsiao WW, Brinkman FS. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics. 2008;9:329.

161. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39:W347–52.

162. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12:402.

163. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002;12:1611–8.

164. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008;24:319–24.

165. Ward N, Moreno-Hagelsieb G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS One. 2014;9:e101850.

166. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WH, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. 2016.

167. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 2012;40:D641–5.

168. Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2014.

169. Popescu A-A, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics. 2012;28:1536–7.

170. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10.

171. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.