PLOS ONE

RESEARCH ARTICLE

# Investigation of horizontal gene transfer of pathogenicity islands in *Escherichia coli* using next-generation sequencing

**Maxim Messerer, Wolfgang Fischer, Sören Schubert***

Max von Pettenkofer-Institut für Hygiene und Medizinische Mikrobiologie, München, Germany

* schubert@med.uni-muenchen.de





**Citation:** Messerer M, Fischer W, Schubert S (2017) Investigation of horizontal gene transfer of pathogenicity islands in *Escherichia coli* using next-generation sequencing. PLoS ONE 12(7): e0179880. https://doi.org/10.1371/journal.pone.0179880

**Editor:** Muna Anjum, Animal and Plant Health Agency, UNITED KINGDOM

**Received:** May 20, 2016

**Accepted:** June 6, 2017

**Published:** July 21, 2017



**Data Availability Statement:** All relevant data are available within the paper and its Supporting Information files. The genome data is available on GenBank/NCBI and the relevant accession numbers are within S1 Table.

**Funding:** This study was supported by a grant by the ERANET "Transnational PathoGenoMics", BMBF.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

Horizontal gene transfer (HGT) contributes to the evolution of bacteria. All extraintestinal pathogenic *Escherichia coli* (ExPEC) harbour pathogenicity islands (PAIs), however relatively little is known about the acquisition of these PAIs. Due to these islands, ExPEC have properties to colonize and invade its hosts efficiently. Even though these PAIs are known to be acquired by HGT, only very few PAIs do carry mobilization and transfer genes required for the transmission by HGT. In this study, we apply for the first time next-generation sequencing (NGS) and *in silico* analyses in combination with *in vitro* experiments to decipher the mechanisms of PAI acquisition in ExPEC. For this, we investigated three neighbouring *E. coli* PAIs, namely the high-pathogenicity island (HPI), the *pks* and the *serU* island. As these PAIs contain no mobilization and transfer genes, they are immobile and dependent on transfer vehicles. By whole genome sequencing of the entire *E. coli* reference (ECOR) collection and by applying a phylogenetic approach we could unambiguously demonstrate that these PAIs are transmitted not only vertically, but also horizontally. Furthermore, we could prove *in silico* that distinct groups of PAIs were transferred "*en bloc*" in conjunction with the neighbouring chromosomal backbone. We traced this PAI transfer *in vitro* using an F' plasmid. Different lengths of transferred DNA were exactly detectable in the sequenced transconjugants indicating NGS as a powerful tool for determination of PAI transfer.

## Introduction

Evolution of bacteria occurs mainly in two major ways, vertical and horizontal. While the vertical transfer is rather slow and inconsistent, the horizontal transfer affects larger parts of the genome and has a greater influence on the evolution of bacteria, especially on the gain of pathogenic properties. Horizontal gene transfer (HGT) can take place by transduction, transformation and conjugation. Plasmids and also larger parts of the genome, like genomic islands, can be conjugated from one bacterium to another [1]. Pathogenicity islands (PAIs) are a subgroup of genomic islands. PAIs encode several virulence factors such as adhesins, toxins, capsules and siderophore systems and play a major role in the evolution of pathogenic bacteria such as extraintestinal pathogenic *Escherichia coli* (ExPEC). ExPECs are responsible for pyelonephritis, cystitis, septicaemia and newborn meningitis [2].

The species *E. coli* is subdivided into four major phylogenetic groups (A, B1, B2 and D). ExPECs belong mostly to groups B2 and D [2]. The *E. coli* reference (ECOR) collection consisting of 72 strains has been shown to represent the genetic diversity of this species. This collection of commensal and pathogenic strains from all phylogenetic groups of *E. coli* was composed in the early 1980s [3].

ExPECs harbour different PAIs, some of which are larger than 100 kb in size [4]. These islands have distinct structural features, e.g. they (i) are integrated at a tRNA gene, (ii) carry a gene for a phage-type integrase and (iii) display a GC-content distinct from that of the chromosomal backbone.

We focused on the high-pathogenicity island (HPI), the *pks* and the *serU* island. These PAIs contribute significantly to the ExPEC virulence [5–7], are next to each other on the chromosome and are not self-transmissible. Therefore, these PAIs are very suitable to investigate large scale HGT within the species *E. coli*.

The HPI is a widespread PAI among *Enterobacteriaceae* and has already been successfully used to demonstrate HGT [5]. This archetypal PAI encodes the synthesis of the siderophore yersiniabactin, representing a highly efficient iron-scavenging molecule. The sequence of the HPI is conserved among different bacterial species, with two distinct types of the island existing in ExPECs: approximately one percent of HPI-positive *E. coli* strains harbour an ICE-type (integrative conjugative element) island, which is completely self-transmissible. About 99% of *E. coli* strains carry a non self-transmissible island with a deletion of about 30 kb, encompassing the mobilization and transfer genes [5].

The two other PAIs, the *serU* island and the *pks* island, carry neither mobilization nor transfer genes [6;7]. The hybrid non-ribosomal peptide-polyketide colibactin encoded by the *pks* island induces double-strand DNA breaks and cell cycle arrest in eukaryotic cells [8]. The virulence factor TcpC encoded by the *serU* island interferes with the innate immune response by interrupting the NF-κB signalling pathway [9]. These two islands are only found in strains of the phylogenetic group B2.

For this study, we sequenced for the first time in large scale the whole genomes of the ECOR collection and some additional strains with next-generation sequencing (NGS). We used two approaches to investigate how HGT contributes to the evolution of PAIs. First, we examined the linked transfer of the described islands and its impact on evolution within the ECOR collection. With NGS, it was possible to compare the phylogeny of the whole PAIs and their neighbouring genomic regions in large scale. Second, we proved the co-transfer of these PAIs with an F' plasmid-mediated conjugation. It was possible to regard potential crossing-over regions for the F' plasmid, to see whether the DNA was conjugated in one or more pieces and to get an overview about the sizes of the transferred DNA.

## Materials and methods

### Bacterial strains, plasmid and primers

The entire 72 strains of the ECOR collection were used as a major set for the NGS approach and the subsequent *in silico* investigation of the phylogeny of *E. coli*. Further *E. coli* strains characterized previously were included in the sequencing project to complement the set of *E. coli* isolates: the strains S107 and S108 from the Le Gall collection [10] reveal a distinct *serU* island [7], the UPEC strain NU14 [11] was successfully used as donor strain in transfer experiments [5]. Finally, we included the archetypal UPEC strain 536 [12]—harboring all three analyzed PAIs—as a reference sequence for the phylogenetic analyses and as donor for the transfer experiments. The K-12 *E. coli* strain MG1655 ($str^R$, phylogenetic group A, no β-hemolysis) [13] was used as recipient strain, as well as its nalidixic acid resistant mutant, which we

**Table 1. Primers.**

| Primer name | Primer sequence |
|---|---|
| ChuA.1 | 5′-GACGAACCAACGGTCAGGAT-3′ |
| ChuA.2 | 5′-TGCCGCCAGTACCAAAGACA-3′ |
| YjaA.1 | 5′-TGAAGTGTCAGGAGACGCTG-3′ |
| YjaA.2 | 5′-ATGGAGAATGCGTTCCTCAAC-3′ |
| TspE4C2.1 | 5′-GAGTAATGTCGGGGCATTCA-3′ |
| TspE4C2.2 | 5′-CGCGCCAACAAAGTATTACG-3′ |
| fyuA.1080.for | 5′-CTACGACATGCCGACAATGCC-3′ |
| fyuA.1709.rev | 5′-TGCTTCCCGCGCCATAACGTG-3′ |
| clbA.IHE.for | 5′-TAACTTCCTTCACTATCTCA-3′ |
| clbA.IHE. rev | 5′-GAGAGGCTAATGCGAGAAAT-3′ |
| tcpC.for | 5′-GGCAACAATATGTATAATATCCT-3′ |
| tcpC.rev | 5'-GCCCAGTCTATTTCTGCTAAAGA-3' |
| HPI-fyuA-2947.rev | 5`-CAACTGCTTCCGTTATAGTGAC-3 |
| HPI-fyuA-2132.for | 5`-AAATTGCGATTAGGACAAATAG-3 |
| p34S-Cm2.484.rev | 5´-TCACCGTAACACGCCACATCTT-3´ |

The Primers which were used in this study are listed. They were used to determine the phylogenetic group (ChuA, YjaA, TspE4C2), to check the presence of the PAIs (*fyuA*, *clbA*, *tcpC*) and the insertion of a chloramphenicol resistance cassette (HPI-*fyuA*, p34S-Cm2).

https://doi.org/10.1371/journal.pone.0179880.t001

constructed for this study. The F' plasmid ($tet^R$) used in this study was isolated from laboratory *E. coli* strain XL1-Blue MRF' (Stratagene; Santa Clara, CA, USA). The primers used in this study are given in Table 1.

## Whole genome sequencing and phylogenetic analysis

The genomic DNA was isolated using the "High Pure PCR Template Preparation Kit" (Roche Diagnostics; Unterhaching, Germany) as indicated by manufacturer's protocol. The sequencing was performed at the Institute Pasteur, Paris (GENOPOLE—Transcriptomics & Epigenomics platform). To construct the libraries, the "TruSeq Kit" (Illumina; San Diego, CA, USA) was used according to manufacturer's instructions. The read type of the HiSeq 2000 (Illumina; San Diego, CA, USA) was single-end 100 nucleotides.

The parameters used for each approach with the NGS data are given in S2 Table.

The raw reads were imported as Illumina data to CLC Genomics Workbench 6.5 (CLC bio; Aarhus, Denmark). After trimming of the sequences, we performed *de novo* assemblies and alignments.

The phylogenetic trees (Maximum Likelihood (ML) with bootstrap and with Bayesian branch support) were constructed using the online tool PhyML 3.0 [14]. The CLC software was also used for the *in silico* MLST applying the Neighbour Joining (NJ) algorithm [15] and to create phylogenetic trees. The trees using ML with Bayesian inference and the analysis of the *E. coli* core genome are present in the manuscript. The trees using NJ and ML with bootstrap are attached to the supplemental section. The selected bootstrap cut-off is 75.

The statistical analysis was performed with the CLC software using the "Create Pairwise Comparison" tool. In order to calculate the DNA homology we used the parameter "percent identity" and to determine the number of Single Nucleotide Polymorphisms (SNPs) the parameter "differences". The software was also used to differentiate between donors and recipients DNA in the genomes of transconjugants of *in vitro* transfer experiments.

The draft genomes were annotated by the PGAP tool from NCBI. The *de novo* assemblies are deposited to NCBI GenBank and the NGS raw reads to NCBI SRA database. The accession numbers are listed in S1 Table.

Beside the PAIs and the neighbouring chromosome, housekeeping gene fragments and the *E. coli* core genome were investigated to definitively determine the phylogenetic groups. We used the Pasteur scheme which includes six housekeeping genes (*trpA*, *trpB*, *pabB*, *putP*, *icd* and *polB*) and has been used for several phylogenetic studies before [16;17].The core genome was analyzed using the tool Parsnp [18]. The closed genome of *E. coli* K-12 strain MG1655 was set as reference. The visualization of the core genome was performed by CLC Genomics Workbench.

## Conjugation and transfer of PAIs

For transfer experiments of the PAIs we used *E. coli* strains NU14 and 536 (phylogenetic group B2, β-hemolysis positive) as donors according to previous protocols [5]. As recipients, the *E. coli* strain MG1655 (phylogenetic group A, β-hemolysis negative) and its nalidixic acid (nal) resistance mutant were used. Dilutions were plated on LB plates containing chloramphenicol (cm) and streptomycin (str) or nalidixic acid to screen for transconjugants. Furthermore, the transconjugants were tested by PCR for the presence of the respective islands as well as the respective phylogenetic group. The β-hemolysis activity was checked on Columbia blood agar plates. The conjugation efficiency (colony forming units (cfu) per ml) was calculated as a ratio between the number of transconjugants and donors [19]. To calculate the efficiency for the transmission of the F' plasmid, we selected tetracycline- (tet) and str-resistant clones. All conjugations were done at least in triplicates for the estimation of efficiency.

$$conjugation\ efficiency = \frac{transconjugants}{donors} cfu/ml$$

# Results

## Simple analysis of draft genomes generated by large scale sequencing

The aim of this study was to decipher large scale horizontal gene transfer (HGT) in *Escherichia coli* affecting its genome. To analyse this, we determined for the first time the draft genomes of all the 72 strains of the *E. coli* reference (ECOR) collection as well as some additional isolates using next-generation sequencing (NGS). The raw data obtained by NGS consisted of a mean number of 9,105,077 reads per genome. A 99.84% of the sequenced nucleotides revealed unambiguous bases and the Phred quality score was 40 on average indicating high quality data [20]. Coverage of the genomes as well as N50 values are given in S1 Table. The *de novo* assembly using the CLC software resulted in about 150 contigs larger than 1 kb with maximum lengths of 145 to 430 kb. To prove the applicability of a phylogenetic approach using draft genomes, we deliberately resigned from performing any additional re-sequencing and gap closure procedures. In order to trace the horizontal transfer and evolution of PAIs, the focus of the present work was on the three neighbouring PAIs, namely the HPI, *pks* and *serU* island as well as the adjacent genomic regions [5;8;9].

With the generated NGS data we performed three approaches. Firstly, the investigation using an *in silico* Multi Locus Sequence Typing (MLST) based on different fragments of housekeeping genes in comparison with the *E. coli* core genome. Secondly, the analysis of the three mentioned neighbouring PAIs and their transmission *in silico*. Thirdly, the transfer of these three PAIs *in vitro* followed by an *in silico* study of the resulting transconjugants.

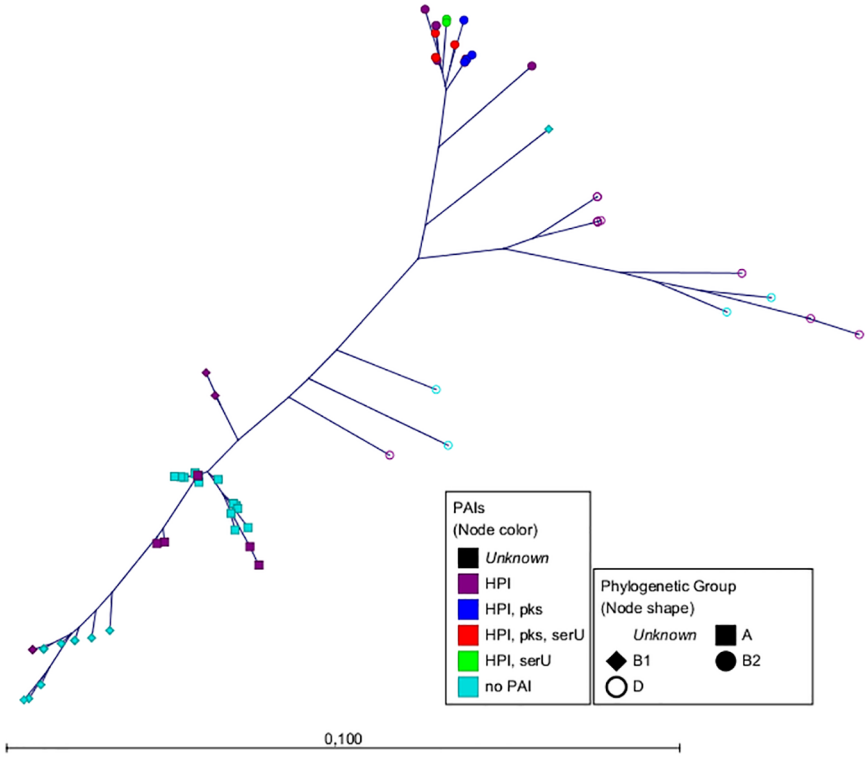


**Fig 1. Radial tree of the six housekeeping gene fragments.** The radial tree of the six housekeeping gene fragments (*trpA*, *trpB*, *pabB*, *putP*, *icd* and *polB*) from the ECOR collection and strains S107, S108 and 536 performed by PhyML using the Maximum Likelihood algorithm with Bayesian branch support. The scale bar represents the number of SNPs per nucleotide. The node colour represents the distribution of the PAIs. The node shapes show the phylogenetic group according to the triplex PCR [2].

https://doi.org/10.1371/journal.pone.0179880.g001

## Investigation of an MLST approach and the core genome to confirm the phylogeny of the *E. coli* species

In order to confirm the ECOR phylogeny and to distinguish between vertical and horizontal PAI transfer in *E. coli* we applied an *in silico* MLST approach and analyzed the species core genome. Several MLST schemes have been described so far to delineate the *E. coli* phylogeny including the Achtman and the Pasteur scheme [21;22]. These schemes rely on the sequence variation of distinct fragments of *E. coli* housekeeping genes with sequence length of about 500 bp to distinguish different phylogenetic groups. The NGS data enabled the comparison of the nucleotide sequences of the gene fragments of the known Pasteur MLST scheme and the *E. coli* core genome to classify the strains. As we investigated the possibility to work with draft genomes without re-sequencing, the tool Parsnp was suitable to analyse the *E. coli* core genome using incomplete genomes.

The fragments of these six housekeeping genes from the Pasteur scheme led to concatenated sequences with a total length of 3,045 bp [16]. To generate an MLST-based tree, we compared the concatenated sequences from all strains using Maximum Likelihood (ML) [14] and the Neighbour-Joining (NJ) algorithm [15]. The constructed phylogenetic tree (Figs 1 and S1) termed "MLST tree" matched highly with previously published data [22–24].

Next, we analyzed the *E. coli* core genome using the tool Parsnp [18] (Fig 2). As reference we set the closed genome of *E. coli* K12-strain (phylogenetic group A). The total coverage among all sequences representing the core genome was 40.9%. This is in total agreement with

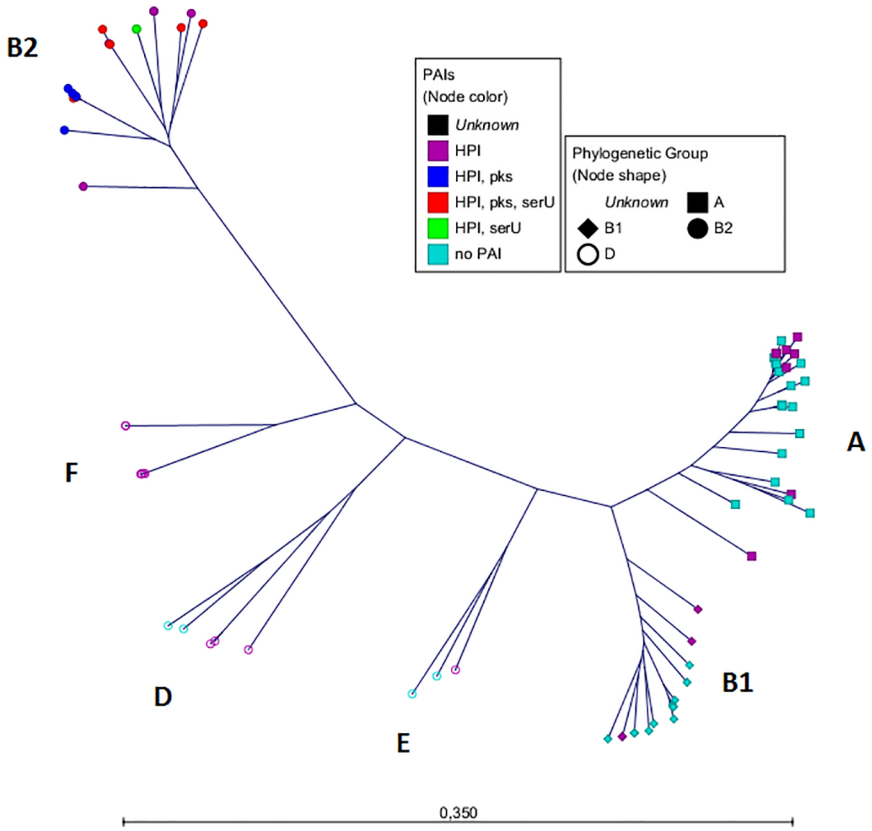


**Fig 2. Radial tree of the *E. coli* core genome.** The radial tree of the core genome was generated by Parsnp. Strain MG1655 was set as reference. The total coverage among all sequences was 40.9%. The phylogenetic groups are highlighted. The scale bar represents the number of SNPs per nucleotide. The node colour represents the distribution of the PAIs. The node shapes show the phylogenetic group according to the triplex PCR [2].

https://doi.org/10.1371/journal.pone.0179880.g002

published data [25;26]. The phylogenetic groups revealed by the analysis of the core genome reflected mostly the groups shown by the "MLST tree". The strains were assigned to the mayor groups A, B1, D and B2 and also to the minor groups E and F [21;22], which are highlighted. The ECOR strain phylogeny is summarized in S3 Table.

## Specific subtypes of the different PAIs are correlated to specific groups

After the investigation of the *E. coli* phylogeny we examined the HGT of immobile PAIs *in silico*. The NGS data enabled a large scale analysis of the HPI, *pks* and *serU* island and their neighbouring chromosomal regions. The genome region under investigation covering all islands and backbone sequences in between encompassed about 126 kb. The draft genome sequences provided sufficient sequence information to analyse this DNA region in all strains.

First, we constructed three phylogenetic trees (NJ, ML with bootstrap and Bayesian inference) comparing the entire HPI (31.5 kb) of all positive strains in order to determine the phylogenetic history of this island (Figs 3 and S2). Interestingly, in all trees we could observe distinct clonal groups of the HPI related to the number and distribution of the neighbouring PAIs. Clonal groups encompassing strains with a distinct number of PAIs were named "PAI-groups". Strains of PAI-group 1 carried only the HPI. PAI-group 2a strains included the HPI and the *pks* island, PAI-group 2b strains harboured the HPI and the *serU* island. In members

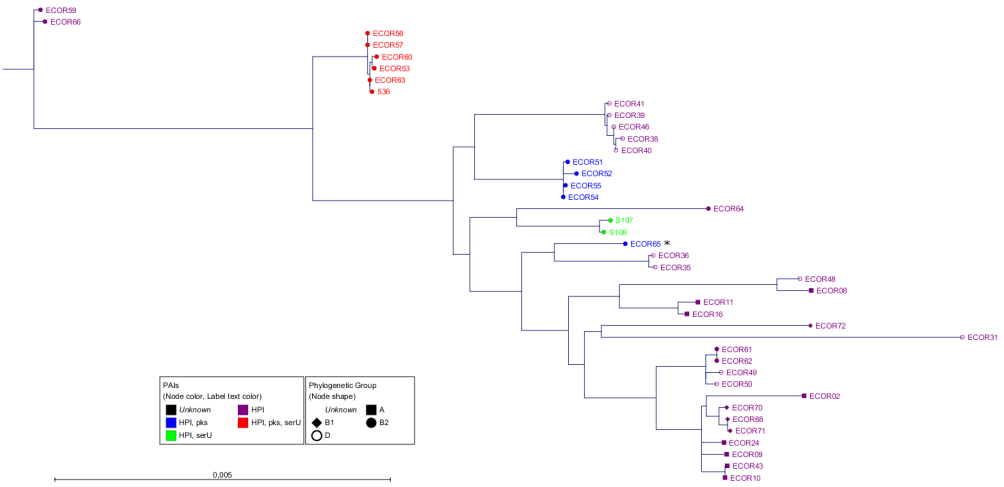


**Fig 3. The phylogenetic tree of the entire HPI.** All strains are at least HPI-positive. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. Except strain ECOR65 (asterisk) from PAI-group 2a, all members of PAI-groups 2a (blue), 2b (green) and 3 (red) showed a HPI subtype specific for their group. The utilized algorithm was Maximum Likelihood with Bayesian branch support performed by PhyML. The scale bar represents the percentage of SNPs per nucleotide. The length of the HPI sequence is about 31.5 kb. The average homology and SNPs within the PAI-groups: 2a 99.99% (4.5), 2b 99.98% (7), 3 99.99% (2.4). The average homology and SNPs between the PAI-groups: 2b-2a 99.63% (116.8), 2b-3 99.54% (144.2), 2a-3 99.59% (129.9). The average homology and SNPs between ECOR65 and the PAI-groups: EC65-2a 99.66% (107.3), EC65-2b 99.67% (104.5), EC65-3 99.53% (149.3).

https://doi.org/10.1371/journal.pone.0179880.g003

of PAI-group 3, all three islands were present. Notably, looking at the phylogram of the HPI (Figs 3 and S2), the formation of PAI-groups 2a and 2b was apparently not due to the deletion of single PAIs from PAI-group 3, as the members of the three different PAI-groups did not share the same clonal group of the HPI. Members of PAI-group 1 revealed different distinct HPI clonal groups as shown in the phylogenetic tree (Fig 3). This heterogeneity suggested that the HPI is the eldest of the three PAIs. Analyses for sequence homology of the entire HPI further corroborated the existence of these clonal groups. The analyses revealed that within the PAI-groups 2a, 2b and 3, the average homology was very high with values between 99.98% and 99.99% (4.5, 7 and 2.4 SNPs, respectively) (Fig 3). In contrast, the homology between different PAI-groups was considerably lower with sequence identities between 99.54% and 99.63% (116.8 to 144.2 SNPs). This indicated that these clonal groups of the HPI had a different phylogenetic history.

Next, the *serU* island (27 kb) and *pks* island (54.5 kb) were each investigated regarding their relationship. The analysis of their genome sequences revealed that the respective *serU* islands (Figs 4 and S3) and *pks* islands (Figs 5 and S4) differed significantly according to the affiliation to the distinct PAI-groups. Each PAI-group consisted of a specific clonal group regarding the respective island. The sequence homology of *serU* islands of strains within PAI-group 2b was 99.93% (18 SNPs). A similar homology was found within PAI-group 3 with 99.94% (16.6 SNPs). However, comparing these two clonal groups, the lower homology of 99.27% (197.3 SNPs) corroborated a different evolution. Regarding the nucleotide sequences of the *pks* islands within PAI-groups 2a, the homology was 99.99% (5.7 SNPs). Within PAI-group 3, a sequence identity of 99.97% (18.3 SNPs) was found. Between these two PAI-groups, the homology was also 99.97% (16 SNPs). Comparing the two different islands, the *pks* islands were in general more similar than the *serU* islands (Figs 4 and 5). This indicated that the *pks* island is probably the most recently acquired of the investigated PAIs.

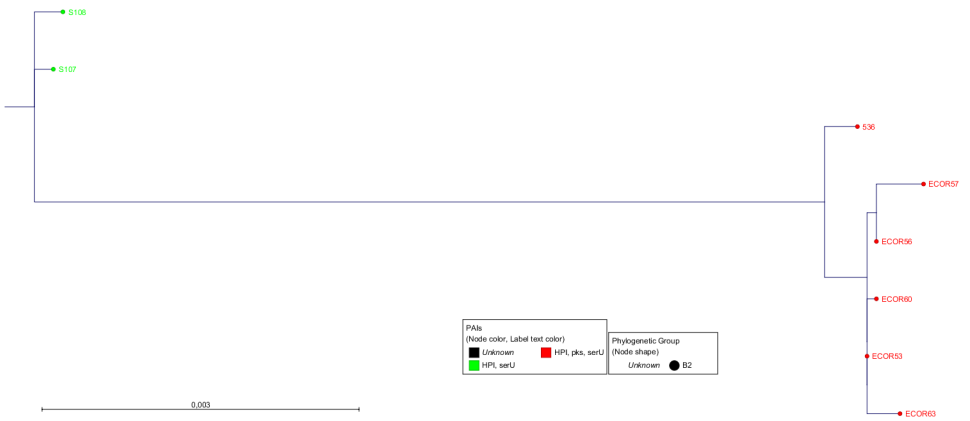


**Fig 4. The phylogenetic tree of the entire *serU* island.** All strains are at least HPI- and *serU* island-positive. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. The members of PAI-groups 2b (green) and 3 (red) showed a *serU* island subtype specific for their group. The algorithm which was used by PhyML was Maximum Likelihood with Bayesian branch support. The scale bar represents the percentage of SNPs per nucleotide. The size of the *serU* island is about 27 kb. The average homology and number of SNPs were similar within the PAI-groups for 2b and 3 with 99.93% (18.0) and 99.94% (16.6), respectively. Between these two PAI-groups, the homology was 99.27% with 197.3 SNPs on average.

https://doi.org/10.1371/journal.pone.0179880.g004

Moreover, we investigated the *E. coli* backbone genome between the islands to gain insight into the diversity of these sequences. If the PAIs were transferred together "*en bloc*", the backbone should cluster and congregate like the PAI subtypes. We named these sequences "inter-PAI regions" which exist between the *serU* island and the HPI (region A, 1 kb) and between the HPI and the *pks* island (region B, 12.5 kb) (Fig 6). We constructed phylogenetic trees (NJ, ML with bootstrap and Bayesian inference) out of these sequences which are shown in Figs 7 and S5 and Figs 8 and S6. These trees resembled that of the HPI-based tree identifying the same PAI-group and indicated that the respective backbone regions were transmitted together

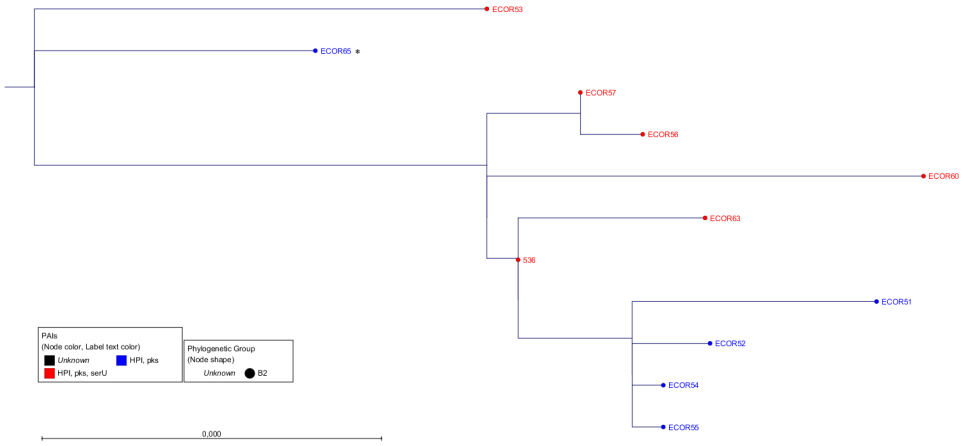


**Fig 5. The phylogenetic tree of the entire *pks* island.** All strains are at least HPI- and *pks* island-positive. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. Except strain ECOR65 (asterisk) from PAI-group 2a, all members of PAI-groups 2a (blue) and 3 (red) showed a *pks* island subtype specific for their group. The algorithm we used was Maximum Likelihood with Bayesian branch support performed by PhyML. The scale bar represents the number of SNPs per nucleotide. The sequence length of the *pks* island is about 54.5 kb. Within PAI-group 2a the homology and the number of SNPS on average are 99.99% and 5.7 respectively, within PAI-group 3 99.97% and 18.3. The average homology and SNPs between ECOR65 and the PAI-groups: EC65-2a 99.89% (61.8), EC65-3 99.89% (59.3).

https://doi.org/10.1371/journal.pone.0179880.g005

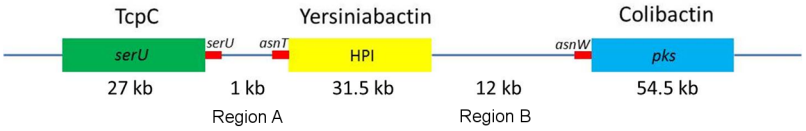


**Fig 6. The arrangement of the PAIs on the chromosome.** Each island is inserted in a tRNA (*serU* island: *serU* tRNA; HPI: *asnT* tRNA; *pks* island: *asnW* tRNA). The size is given in kilobases (kb). The regions between the PAIs are called inter-PAI regions (between the *serU* island and the HPI (region A): about 1 kb; between the HPI and the *pks* island (region B): about 12 kb).

https://doi.org/10.1371/journal.pone.0179880.g006

"*en bloc*" with the islands. The analyses of sequence homologies corroborated this hypothesis (Figs 7 and 8). The sequences of region A were almost identical within the PAI-groups 2a, 2b and 3 with sequence identities from 99.97% to 100%, whereas those between the PAI-groups were definitely lower (99.60% to 99.78%). Additionally, the sequences of region B within respective PAI-groups exhibited very high homologies (99.98% to 99.99%). These sequence homologies were less pronounced between the three PAI-groups (99.28% to 99.87%).

As third level of evidence, we wanted to ensure that the PAIs were transferred via HGT and not vertically through cell division. For this purpose we compared the phylogenetic tree of the core genome (Fig 2) with that of the PAIs (Figs 3–5). If the sequences of the core genome cluster together, the strains originated from the same ancestor. In contrast, a sequence variety indicates a different origin and supports the idea of a transmission of the PAIs *via* HGT. The fact that strains with the same PAI distribution did not seem to be clonal regarding their core genome sequences (Fig 2) proved that the transmission of the PAIs was not vertical, but horizontal.

## Analysis of the outlier strain ECOR65 which showed a distinct phylogenetic pattern

By analysing the ECOR collection *in silico*, we found one strain, namely ECOR65, which did not fit into the proposed scheme. ECOR65 harboured the PAIs HPI and *pks* island and was

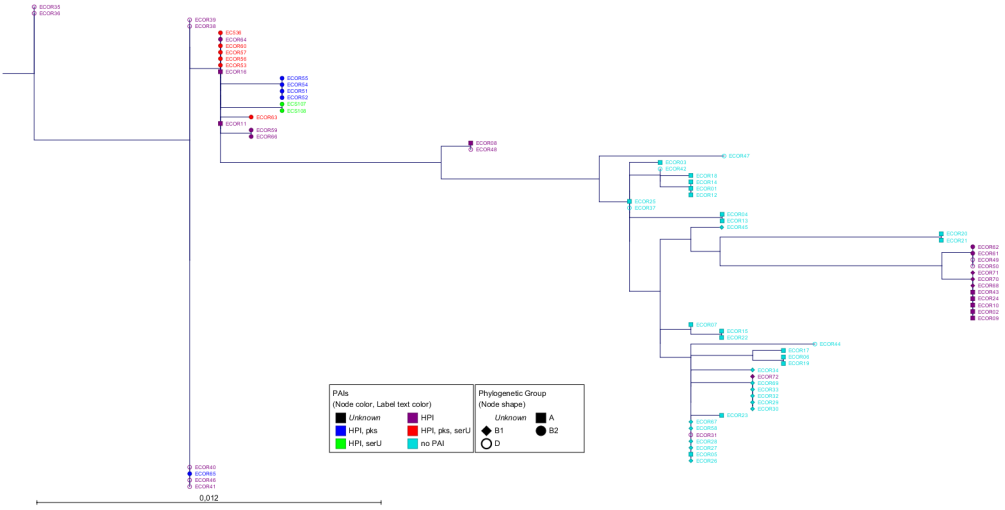


**Fig 7. Phylogenetic tree of region A.** The inter-PAI region between the *serU* island and the HPI (region A) is about 1 kb and is shown as phylogenetic tree. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. The algorithm which was used by PhyML was Maximum Likelihood with Bayesian branch support. The scale bar represents the percentage of SNPs per nucleotide. Within PAI-groups 2a and 2b, the percental homology is 100% without any SNP. Within PAI-group 3, the homology is 99.97% and the number of SNPs 0.3 on average.

https://doi.org/10.1371/journal.pone.0179880.g007

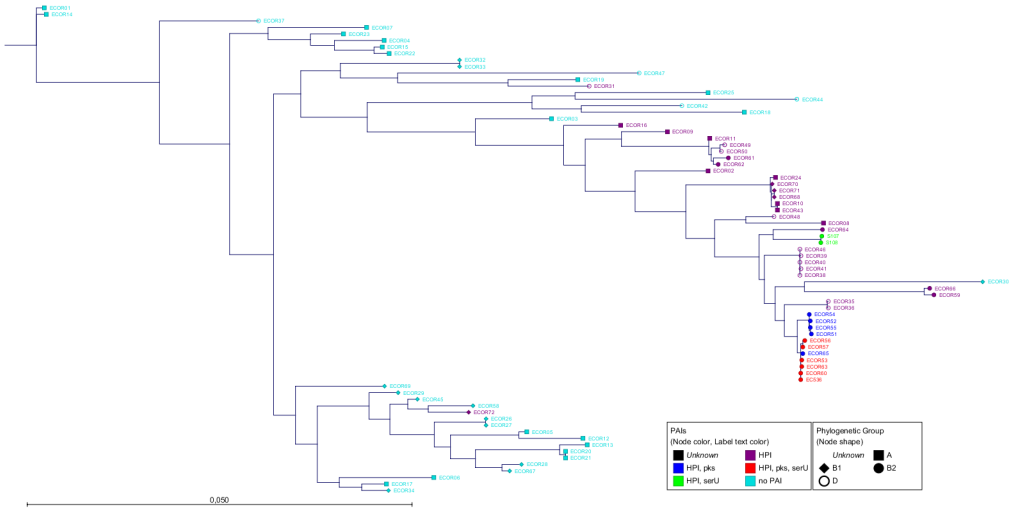


**Fig 8. Phylogenetic tree of region B.** The dendrogram of the inter-PAI region between the HPI and the *pks* island (region B). The size of the sequence is about 12 kb. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. The algorithm which was used by PhyML was Maximum Likelihood with Bayesian branch support. The scale bar represents the number of SNPs per nucleotide. For the PAI groups 2a, 2b and 3, the average homology was 99.99%,99.98% and 99.98% respectively and the number of SNPs were 1.3, 3 and 2.3 on average.

https://doi.org/10.1371/journal.pone.0179880.g008

thus, by definition, a member of PAI-group 2a. Comparing the HPI sequence of ECOR65 with those of the PAI-groups 2a, 2b and 3, the HPI of ECOR65 was only distantly related (Fig 3). In contrast, looking at the phylogenetic trees of the *pks* island (Fig 5) and the inter-PAI regions (Figs 7 and 8), the strain resembled members of PAI-group 3, suggesting the loss of the *serU* island in ECOR65. To prove whether ECOR65 lost this island we divided the inter-PAI regions into equal parts and analysed them separately to examine the distribution of the genetic differences. The region between the *serU* island and the HPI (region A) is 1 kb in length. By investigating the two 500 bp parts of region A we found the part next to the HPI to cluster ECOR65 together with strains of PAI-group 3. In the other 500 bp part the sequence of ECOR65 had an aberration of only one single SNP compared to the sequences of the other strains, which we regarded as non-discriminatory. The length of the inter-PAI region between the HPI and the *pks* island (region B) is about 12.5 kb. We divided this region into equal sized parts and analysed the respective phylogenetic trees. The dendrogram of the 6 kb region next to the HPI classified ECOR65 strain to be of PAI-group 3. In contrast, the tree of the region next to the *pks* island clustered strains of PAI-group 3 and 2a together with ECOR65. However, the latter 6 kb region was regarded as non-discriminatory displaying only an average of 1.4 SNPs. In conclusion, these data pointed towards the loss of the *serU* island in the ECOR65 strain, but did not explain the sequence difference of HPI of ECOR65 to members of the PAI-group 3.

## PAIs are transmissible *via* F' plasmid-mediated transfer

After the *in silico* investigation of the ECOR collection and distinct additional strains, the hypothesis of an "*en bloc*" transfer was proven applying an *in vitro* approach. We performed an F' plasmid-mediated conjugation to reconstruct the simultaneous transmission of multiple PAIs. The resulting transconjugants were sequenced and further analyzed. We focused (i) on the amount of transferred DNA from different donor strains and (ii) on the recognition of potential hotspots for recombination. NGS is a powerful tool to retrace an "*en bloc*" transfer and to gain insight into the evolution of the PAIs and the surrounding backbone. The sequenced

genomes of donors, transconjugants and recipients were compared by an alignment to identify the exact regions of homologous recombination.

We used both *E. coli* strain NU14 HPI-Cm F' carrying the HPI together with the *pks* island (PAI-group 2a) and *E. coli* 536 HPI-Cm F' harbouring all three islands (PAI-group 3) as donor strains to confirm our hypothesis. The respective HPIs were tagged with a chloramphenicol resistance cassette applying the method of Datsenko and Wanner [27] to track the transfer of the PAIs. As the investigated islands were immobile, an F' plasmid was conjugated to the donors to enable the transfer. As the transmissions were conducted to the recipient *E. coli* MG1655, the further characteristics of selected transconjugants were to belong to phylogenetic group A and to reveal no β-hemolysis on blood agar plates. In order to compare the conjugation efficiency of an F' plasmid transfer in general ($str^R$, $tet^R$) with a PAI transmission in special ($str^R$, $cm^R$), we calculated the ratio between the number of transconjugants and donors [19]. The transconjugants were checked by PCR for the presence of the HPI and the *pks* island, and in case of the donor strain 536 HPI-Cm F', additionally for the presence of the *serU* island.

The pure F' plasmid transfer exhibited a conjugation efficiency of $5.24 \times 10^{-4}$ cfu/ml. None of 120 screened clones were resistant to chloramphenicol indicating a transmission of other DNA content than the HPI. In contrast to this conjugation, the HPI transfer rate of donor NU14 HPI-Cm F' (HPI and *pks* island) was significantly lower with $2.94 \times 10^{-7}$ cfu/ml. The conjugation efficiency for the HPI transfer of donor 536 HPI-Cm F' (HPI, *pks* and *serU* island) was very similar with $3.85 \times 10^{-7}$ cfu/ml. This indicated that the efficiency was independent of the donor. Interestingly, 50% of the transconjugants were tetracycline-resistant, indicating the retention of the F' plasmid. After three passages without antibiotic pressure 90% of the initially tetracycline-resistant strains were not able to grow on tetracycline-LB-plates any more. The loss of the resistance could reflect the recombination *via* double crossing-over (exchange of donor and recipient DNA) and the loss of the F' plasmid. Notably, the tetracycline-resistant transconjugants were able to spread their new PAIs with a 100-fold enhanced conjugation efficiency of $4.14 \times 10^{-5}$ cfu/ml. This transfer rate resembled more the pure F' plasmid transfer than the PAI transfer.

Next, we analysed the transconjugants *in silico*. For this, we sequenced the whole genomes of the respective transconjugants to gain insight into the F' plasmid transfer of the PAIs, their backbone and the recombination into the recipients. We took tetracycline-sensitive transconjugants from five independent conjugations of donor NU14 HPI-Cm F' (PAI-group 2a) and from four independent conjugations of donor 536 HPI-Cm F' (PAI-group 3). The sequences of all investigated transconjugants revealed exclusively unfragmented DNA transfer events, with only one piece of foreign DNA found per isolate. Furthermore, the tetracycline-negative strains revealed no F' plasmid DNA in their genome indicating a double, rather than a single, crossing-over. The PAIs were always transmitted completely from donors to recipients. This could be due to the fact that no homologous regions (related to the islands) were present in the recipient. Also no IS elements contributing to recombination are described in the three PAIs.

With the entire genome sequences of donor and recipient strains in our hands, the NGS approach enabled us to distinguish between DNA of donor and recipient within the transconjugant sequences. The comparison of donors, recipients and transconjugants showed that the size of integrated DNA was highly variable (Table 2). The transferred DNA from the PAI-group 2a strain NU14 HPI-Cm F' varied between 131,132 bp and 421,058 bp. The PAI-group 3 strain 536 HPI-Cm F' transferred DNA fragments from 62,496 bp to 470,591 bp in size. This indicated that the size of integrated DNA was independent of the donor. Mostly, no regular hotspots for recombination of the F' plasmid within the chromosome were detectable. Although one integration site was similar in three transconjugants, the recombination took place at various locations within the genome. All sites were analysed for IS elements in the

**Table 2. Transconjugants.**

| Transconjugant | size of donor DNA [bp] |
|---|---|
| NU14 HPI-Cm F' x MG1655 K1 | 223,368–224,149 |
| NU14 HPI-Cm F' x MG1655 K2 | 321,955–322,073 |
| NU14 HPI-Cm F' x MG1655 K3 | 419,923–421,058 |
| NU14 HPI-Cm F' x MG1655 K4 | 131,132–132,269 |
| NU14 HPI-Cm F' x MG1655 K5 (no *pks* island) | 198,334–198,400 |
| 536 HPI-Cm F' x MG1655 K1 | 225,011–226,036 |
| 536 HPI-Cm F' x MG1655 K2 | 470,189–470,591 |
| 536 HPI-Cm F' x MG1655 K3 | 348,167–348,408 |
| 536 HPI-Cm F' x MG1655 K4 (no *pks* island) | 62,496–62,711 |

The transconjugants from the independent conjugations NU14 HPI-Cm F' x MG1655 and 536 HPI-Cm F' x MG1655 are listed. The amount of transferred donor DNA is given in base pairs (bp).

https://doi.org/10.1371/journal.pone.0179880.t002

genome of *E. coli* MG1655, but none were found at these places. This finding could relate to the fact that recombination of an F' plasmid into the chromosome can occur at any homologous region [28].

Interestingly, only 64 of 70 transconjugants (91.4%) analysed by PCR had obtained the *pks* island from the donors. No residual sequences of the *pks* island were found *in silico* in the two sequenced isolates which had previously been identified as PCR-negative. This indicated a transfer in an "all or nothing" fashion.

## Discussion

The 72 strains of the *Escherichia coli* reference (ECOR) collection were composed in the early 1980s from a selection of 2,600 *E. coli* isolates to represent the range of genotypic variation in the species as a whole [3]. Although the ECOR collection does not fully represent the different pathotypes of *E. coli*, it is suitable to analyse the horizontal gene transfer (HGT) of extraintestinal pathogenic *E. coli* (ExPEC) pathogenicity islands (PAIs) due to the presence of this pathotype in the compilation [24]. This strain collection was used for a variety of publications over the last three decades. Differences in some strains have been described within the ECOR collection, mainly regarding their published virulence factors [29]. Therefore, next-generation sequencing (NGS) is a powerful tool enabling the definite verification of the respective ECOR strains and delivering the adequate sequence data for phylogenetic comparisons of the strains. Meanwhile, the cost for whole genome sequencing decreased enormously compared to the introduction of this new technology [30]. Besides the ECOR collection, we used three archetypal ExPEC strains for our study: strain 536 [12], S107 and S108 [10]. As ExPECs are relevant clinical pathogens with virulence often linked to PAIs [31], they are ideal to investigate PAI transfer in *E. coli*.

To our knowledge, the HGT of large DNA regions has not been studied by an NGS approach in such a comprehensive strain collection. Due to the high quality raw data obtained by NGS, the assembly of draft genomes containing large contigs was possible. The aim of this study was to analyse if these NGS draft genomes are sufficient for (i) phylogenetic analyses of the core genome and (ii) defining PAI transfer in *E. coli* by combining *in silico* and *in vitro* experiments. In order to determine the phylogenetic groups of the ECOR collection and the additional strains, we analyzed the *E. coli* core genome and additionally applied an *in silico* MLST by analysing the housekeeping genes of the Pasteur MLST scheme (*trpA*, *trpB*, *pabB*, *putP*, *icd* and *polB*), which seemed to be only little affected by HGT and recombination events

[16;32]. We constructed two phylogenetic trees using the standard 500 bp fragments of these six genes ("MLST tree") and the core genome. The structure of the "MLST tree" was in almost perfect agreement with the respective data of previous studies [22–24].

MLST is a good approach for rough classification of strains in phylogenetic groups, but by NGS the analysis of the core genome and the yield of higher resolution of phylogeny is possible. As the draft genomes of the entire ECOR collection are now available for the scientific community, extensive *in silico* approaches can be performed to investigate genomic regions of interest. Furthermore, the draft genomes were used to analyse the DNA acquisition combining *in silico* and *in vitro* experiments with a focus on three PAIs, namely the HPI, the *pks* and the *serU* island as well as the surrounding backbone genome. Due to the distribution of the three PAIs, we classified the analysed strains into distinct clusters named PAI-groups. We were able to demonstrate, that the isolates of the different PAI-groups carried distinct PAI subtypes regarding the respective islands. Also the neighbouring backbone regions were nearly identical within each PAI-group. In contrast, the housekeeping genes scattered around the genome showed no relationship among the strains within each group. This evidenced the same origin of these PAIs and the directly adjacent backbone genome underscoring their horizontal "*en bloc*" transfer.

Our 3-way comparison of phylogenetic trees using ML with bootstrap, ML with Bayesian inference and NJ showed differences in clustering ECOR strains, but the conclusions to the horizontal PAI transfer was supported by all three methods.

In order to reproduce the "*en bloc*" transfer indicated by our *in silico* data, we constructed two donors: one harbouring the HPI and the *pks* island (NU14 HPI-Cm F'; PAI-group 2a) and one harbouring all three PAIs (536 HPI-Cm F'; PAI-group 3). To apply a transfer of immobile PAIs, we used an F' plasmid which transfers the donors' DNA to the recipients leading to integration. The resulting transconjugants were fully sequenced and the NGS data enabled us to investigate exactly the size of the integrated donor DNA. We could transfer up to 470.5 kb, which is almost 10% of the whole *E. coli* genome. These transfer events showed that in most cases all PAIs were transmitted together with the directly adjacent backbone. Interestingly, the *pks* island was only transferred into 91.4% of transconjugants from PAI-group 3. This indicated that the PAIs of the different PAI-groups were not always completely transferred. In contrast, we never observed partial transfer of the islands leading to fragmented PAIs. This was probably due to the required sequence homology of donor's and recipient's DNA, suggesting an "all or nothing" transfer.

Of note, the tetracycline-resistant transconjugants were able to further spread the received PAIs with higher conjugation efficiency. This could be the reason for the broad distribution of the *E. coli* HPI although this PAI is not self-transferable. Interestingly, the ICE-type of the HPI, which is still mobile, is less present in the *E. coli* species [5;33].

In the phylogenetic analyses of the *pks* island and the inter-PAI regions, the strain ECOR65 from PAI-group 2a did not cluster together with strains of the respective PAI-group, but could be assigned to PAI-group 3 strains. Instead, the HPI sequence resembled those of PAI-group 1, which were scattered over the phylogenetic tree of the HPI. One possible explanation would be that ECOR65 gained all three PAIs and the neighbouring backbone from a member of PAI-group 3 and subsequently lost the *serU* island in a deletion event. It was reported that the region surrounding the *serU* island and the HPI is a hotspot of recombination [34] and a "bastion of polymorphism" [34;35]. Nevertheless, this hypothesis doesn't explain the divergence between the ECOR65-HPI and the HPIs of PAI-group 3 strains. Another explanation is based on our findings of the *in vitro* approach. According to the phylogenetic tree of the HPI (Fig 3), ECOR65 was a former PAI-group 1 isolate. Then, a PAI-group 3 strain might have transferred only the *pks* island instead of three PAIs, representing an incomplete transfer event of the

respective islands of the PAI group. This would lead to a clonal HPI subtype of PAI-group 1 and a clonal *pks* island subtype with surrounding backbone genome of PAI-group 3. This hypothesis is supported by the fact that especially the inter-PAI region between the HPI and the *pks* island were identical (Fig 8).

For the first time, we were able to reconstruct the HGT of large genomic regions by a combination of *in silico* and *in vitro* experiments due to NGS. The results showed that the exchange of immobile *E. coli* PAIs between *E. coli* isolates also influences the genomic backbone. Further data have to follow to completely understand the dimension of this transfer.

## Supporting information

**S1 Fig. Radial tree of the six housekeeping gene fragments.** The radial tree of the six housekeeping gene fragments (*trpA*, *trpB*, *pabB*, *putP*, *icd* and *polB*) from the ECOR collection and strains S107, S108 and 536. The scale bar represents the number of SNPs per nucleotide. The node colour represents the distribution of the PAIs. The node shapes show the phylogenetic group according to the triplex PCR [2]. a) Tree performed by PhyML using the Maximum Likelihood algorithm with bootstrap. b) Tree performed by CLC Genomics Workbench using the Neighbour-Joining algorithm.
(TIF)

**S2 Fig. The phylogenetic tree of the entire HPI.** All strains are at least HPI-positive. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. Except strain ECOR65 (asterisk) from PAI-group 2a, all members of PAI-groups 2a (blue), 2b (green) and 3 (red) showed a HPI subtype specific for their group. The scale bar represents the percentage of SNPs per nucleotide. a) The utilized algorithm was Maximum Likelihood with bootstrap performed by PhyML. b) The utilized algorithm was Neighbour-Joining performed by CLC Genomics Workbench.
(TIF)

**S3 Fig. The phylogenetic tree of the entire *serU* island.** All strains are at least HPI- and *serU* island-positive. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. The members of PAI-groups 2b (green) and 3 (red) showed a *serU* island subtype specific for their group. The scale bar represents the percentage of SNPs per nucleotide. a) The algorithm which was used by PhyML was Maximum Likelihood with bootstrap. b) The algorithm which was used by CLC Genomics Workbench was Neighbour-Joining.
(TIF)

**S4 Fig. The phylogenetic tree of the entire *pks* island.** All strains are at least HPI- and *pks* island-positive. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. Except strain ECOR65 (asterisk) from PAI-group 2a, all members of PAI-groups 2a (blue) and 3 (red) showed a *pks* island subtype specific for their group. The scale bar represents the number of SNPs per nucleotide. a) The algorithm we used was Maximum Likelihood with bootstrap performed by PhyML. b) The algorithm we used was Neighbour-Joining performed by CLC Genomics Workbench.
(TIF)

**S5 Fig. Phylogenetic tree of region A.** The inter-PAI region between the *serU* island and the HPI (region A) is shown as phylogenetic tree. The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. The scale bar represents the percentage of SNPs per nucleotide. a) The algorithm which was used by PhyML was Maximum Likelihood with bootstrap. b) The algorithm which was used by CLC Genomics Workbench was

Neighbour-Joining.
(TIF)

**S6 Fig. Phylogenetic tree of region B.** The dendrogram of the inter-PAI region between the HPI and the *pks* island (region B). The text and dot colour represents the PAI-group and the dot shape the phylogenetic group. The scale bar represents the number of SNPs per nucleotide. a) The algorithm which was used by PhyML was Maximum Likelihood with bootstrap. b) The algorithm which was used by CLC Genomics Workbench was Neighbour-Joining.
(TIF)

**S1 Table. NGS accessions.**
(XLSX)

**S2 Table. NGS parameters.**
(DOCX)

**S3 Table. ECOR phylogeny.**
(XLSX)

## Acknowledgments

We thank Phillippe Glaser and his team for the sequencing of the *E. coli* isolates. This study was supported by a grant by the ERANET "Transnational PathoGenoMics", BMBF.

## Author Contributions

**Conceptualization:** SS WF.

**Data curation:** MM.

**Formal analysis:** MM WF.

**Funding acquisition:** SS.

**Investigation:** MM.

**Methodology:** SS WF MM.

**Project administration:** SS WF.

**Resources:** SS.

**Software:** MM.

**Supervision:** SS WF.

**Validation:** SS WF MM.

**Visualization:** SS WF.

**Writing – original draft:** MM SS.

**Writing – review & editing:** SS WF MM.

## References

1. Schneider G, Dobrindt U, Middendorf B, Hochhut B, Szijarto V, Emody L, et al. Mobilisation and remobilisation of a large archetypal pathogenicity island of uropathogenic *Escherichia coli* in vitro support the role of conjugation for horizontal transfer of genomic islands. BMC Microbiol 2011; 11:210. https://doi.org/10.1186/1471-2180-11-210 PMID: 21943043

2. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl Environ Microbiol 2000 Oct; 66(10):4555–8. PMID: 11010916
3. Ochman H, Selander RK. Standard reference strains of *Escherichia coli* from natural populations. J Bacteriol 1984 Feb; 157(2):690–3. PMID: 6363394
4. Dobrindt U, Blum-Oehler G, Nagy G, Schneider G, Johann A, Gottschalk G, et al. Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536. Infect Immun 2002 Nov; 70(11):6365–72. https://doi.org/10.1128/IAI.70.11.6365-6372.2002 PMID: 12379716
5. Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, Hoffmann C, et al. Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. PLoS Pathog 2009 Jan; 5(1):e1000257. https://doi.org/10.1371/journal.ppat.1000257 PMID: 19132082
6. Putze J, Hennequin C, Nougayrede JP, Zhang W, Homburg S, Karch H, et al. Genetic structure and distribution of the colibactin genomic island among members of the family *Enterobacteriaceae*. Infect Immun 2009 Nov; 77(11):4696–703. https://doi.org/10.1128/IAI.00522-09 PMID: 19720753
7. Schubert S, Norenberg D, Clermont O, Magistro G, Wieser A, Romann E, et al. Prevalence and phylogenetic history of the TcpC virulence determinant in *Escherichia coli*. Int J Med Microbiol 2010 Nov; 300 (7):429–34. https://doi.org/10.1016/j.ijmm.2010.02.006 PMID: 20547102
8. Nougayrede JP, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, et al. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science 2006 Aug 11; 313(5788):848–51. https://doi.org/10.1126/science.1127059 PMID: 16902142
9. Cirl C, Wieser A, Yadav M, Duerr S, Schubert S, Fischer H, et al. Subversion of Toll-like receptor signaling by a unique family of bacterial Toll/interleukin-1 receptor domain-containing proteins. Nat Med 2008 Apr; 14(4):399–406. https://doi.org/10.1038/nm1734 PMID: 18327267
10. Le GT, 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 Nov; 24(11):2373–84. https://doi.org/10.1093/molbev/msm172 PMID: 17709333
11. Johnson JR, Weissman SJ, Stell AL, Trintchina E, Dykhuizen DE, Sokurenko EV. Clonal and pathotypic analysis of archetypal *Escherichia coli* cystitis isolate NU14. J Infect Dis 2001 Dec 15; 184(12):1556–65. https://doi.org/10.1086/323891 PMID: 11740731
12. Hacker J, Ott M, Blum G, Marre R, Heesemann J, Tschape H, et al. Genetics of *Escherichia coli* uropathogenicity: analysis of the O6:K15:H31 isolate 536. Zentralbl Bakteriol 1992 Jan; 276(2):165–75. PMID: 1559005
13. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997 Sep 5; 277(5331):1453–62. PMID: 9278503
14. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010 May; 59(3):307–21. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638
15. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987 Jul; 4(4):406–25. PMID: 3447015
16. Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le BC, Denamur E. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. Mol Biol Evol 2004 Jun; 21(6):1085–94. https://doi.org/10.1093/molbev/msh118 PMID: 15014151
17. 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 Oct; 10 (10):2484–96. https://doi.org/10.1111/j.1462-2920.2008.01669.x PMID: 18518895
18. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 2014; 15(11):524. https://doi.org/10.1186/s13059-014-0524-x PMID: 25410596
19. Beaber JW, Hochhut B, Waldor MK. SOS response promotes horizontal dissemination of antibiotic resistance genes. Nature 2004 Jan 1; 427(6969):72–4. https://doi.org/10.1038/nature02241 PMID: 14688795
20. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998 Mar; 8(3):186–94. PMID: 9521922
21. 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 Jun; 60(5):1136–51. https://doi.org/10.1111/j.1365-2958.2006.05172.x PMID: 16689791
22. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, et al. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics 2008; 9:560. https://doi.org/10.1186/1471-2164-9-560 PMID: 19036134

23. Martin P, Marcq I, Magistro G, Penary M, Garcie C, Payros D, et al. Interplay between siderophores and colibactin genotoxin biosynthetic pathways in *Escherichia coli*. PLoS Pathog 2013; 9(7):e1003437. https://doi.org/10.1371/journal.ppat.1003437 PMID: 23853582

24. Chaudhuri RR, Henderson IR. The evolution of the *Escherichia coli* phylogeny. Infect Genet Evol 2012 Mar; 12(2):214–26. https://doi.org/10.1016/j.meegid.2012.01.005 PMID: 22266241

25. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 2010 Nov; 60(4):708–20. https://doi.org/10.1007/s00248-010-9717-3 PMID: 20623278

26. Gordienko EN, Kazanov MD, Gelfand MS. Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*4. J Bacteriol 2013 Jun; 195(12):2786–92. https://doi.org/10.1128/JB.02285-12 PMID: 23585535

27. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A 2000 Jun 6; 97(12):6640–5. https://doi.org/10.1073/pnas.120163297 PMID: 10829079

28. Griffiths A, Gelbart W, Miller J, Lewontin R. Modern Genetic Analysis. 1999.

29. Johnson JR, Delavari P, Kuskowski M, Stell AL. Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*. J Infect Dis 2001 Jan 1; 183(1):78–88. https://doi.org/10.1086/317656 PMID: 11106538

30. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 2012 Apr; 50(4):1355–61. https://doi.org/10.1128/JCM.06094-11 PMID: 22238442

31. Kohler CD, Dobrindt U. What defines extraintestinal pathogenic *Escherichia coli*? Int J Med Microbiol 2011 Dec; 301(8):642–7. https://doi.org/10.1016/j.ijmm.2011.09.006 PMID: 21982038

32. Lecointre G, Rachdi L, Darlu P, Denamur E. *Escherichia coli* molecular phylogeny using the incongruence length difference test. Mol Biol Evol 1998 Dec; 15(12):1685–95. PMID: 9866203

33. Schubert S, Dufke S, Sorsa J, Heesemann J. A novel integrative and conjugative element (ICE) of *Escherichia coli*: the putative progenitor of the *Yersinia* high-pathogenicity island. Mol Microbiol 2004 Feb; 51(3):837–48. PMID: 14731283

34. 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 Jan; 5(1):e1000344. https://doi.org/10.1371/journal.pgen.1000344 PMID: 19165319

35. Milkman R. Recombination and population structure in *Escherichia coli*. Genetics 1997 Jul; 146(3):745–50. PMID: 9215884