Supplementary Online Material

Supplementary Online Text

Lateral gene transfer versus groups and/or virulence.

Within the ECOR collection, the copy numbers of the insertion elements IS1 through IS5 and IS30 distinguish A (22.7 copies ± 2.5) and D (virulent, 21.6 ± 4.8) from the virulent groups B2 (8.2 ± 1.0) and B1 (6.3 ± 2.0) (t-test, $P < 0.01$) (Sawyer et al. 1987). Additional analyses of the numbers of transposases and transposase-like sequences in published *E. coli* genomes also show no clear correlation with virulence or group, except that transposons are particularly frequent in *Shigella flexneri* (ABD, strain 2457T: 152.9 copies/MB; strain 301: 66.2) due to "local hopping" (Boyd and Hartl 1997) in contrast to K12 (A, 20.7), the uropathogen CFT073 (B2, 18.7) or O157:H7 strains (ABD, 9.2). Thus, there does not seem to be a clear association with the numbers of IS elements or transposases and virulence.

One might expect more frequent recombination among virulent lineages to be an indirect consequence of a higher frequency of genetic exchange by transduction and conjugation. Among other factors, the efficiency of conjugation and transduction is reduced by restriction enzymes that degrade unmodified DNA. Using the same strategy as above, we screened the *E. coli* genomes for coding sequences (CDSs) encoding restriction/modification enzymes. Again, no clear pattern could be discerned. Both of the virulent *Shigella* genomes possess nearly one hundred CDSs encoding restriction/modification genes but K12 does as well, whereas the O157:H7 and CFT073 genomes possess < 10 copies. Thus, it seems unlikely that more frequent homologous recombination among virulent lineages reflects increased opportunities for genetic exchange.

Table S1. Oligonucleotide primers used for *E. coli* MLST

Locus size	Gene	Primer	Sequence (5′ – 3′)
adk (536)	adenylate kinase	adk-P1	ATTCTGCTTGGCGCTCCGGG
		adk-P2	CCGTCAACTTTCGCGTATTT
fumC (469)	fumarate hydratase	fumC-P1	TCACAGGTCGCCAGCGCTTC
		fumC-P2	GTACGCAGCGAAAAAGATTC
gyrB (460)	DNA gyrase	gyrB-P1	TCGGCGACACGGATGACGGC
		gyrB-P2	ATCAGGCCTTCACGCGCATC
icd (518)	isocitrate/isopropylmalate dehydrogenase	icd-P1	ATGGAAAGTAAAGTAGTTGTTCCGGCACA
		icd-P2	GGACGCAGCAGGATCTGTT
mdh (452)	malate dehydrogenase	mdh-P1	ATGAAAGTCGCAGTCCTCGGCGCTGCTGGCGG
		mdh-P2	TTAACGAACTCCTGCCCCAGAGCGATATCTTTCTT
purA (478)	adenylosuccinate dehydrogenase	purA-P1	CGCGCTGATGAAAGAGATGA
		purA-P2	CATACGGTAAGCCACGCAGA
recA (510)	ATP/GTP binding motif	recA-P1	CGCATTCGCTTTACCCTGACC
		recA-P2	TCTCGATCAGCTTCTCTTTT

Table S2. Sources of 460 *E. coli* by group

Category	Subdivision (Total)	A	AxB1	B1	B2	D	ABD
Host:	domesticated (61)	16	6	12	15	5	7
	captive (26)	4	5	6	6	3	2
	wild (24)	1	3	8	5	5	2
	human (346)	78	48	82	43	20	75
Continent	Europe (170)	19	10	43	50	16	32
	Africa (118)	62	22	20	0	6	8
	North America (58)	11	9	11	12	3	12
	Asia+Pacific (24)	4	6	4	4	4	2
Pathogens	K1 (46)	0	1	0	30	3	12
	EHEC (41)	1	4	23	0	1	12
	EIEC (38)	0	13	12	0	2	11
	EPEC (20)	1	0	9	6	0	4
	ETEC (13)	7	3	2	1	0	0
	APEC (13)	0	0	4	5	1	3
	EAEC (9)	5	0	1	0	3	0
	S. dysenteriae (23)	0	1	9	0	4	9
	S. boydii (19)	0	1	9	0	0	9
	S. flexneri (16)	0	1	2	0	0	13
	S. sonnei (3)	0	0	3	0	0	0

Table S3: Nucleotide diversity (π), Tajima's D test and Homoplasy (*H*) ratio test for the four *E. coli* populations, as well as for the two major hybrid groups. The top half of the table includes all isolates and the bottom half, STs only.

Population	Sample size	π	Tajima's D	*H* ratio
A	99	0.004	-0.882	
B1	106	0.006	-0.102	
B2	70	0.005	-1.084	
D	34	0.013	-0.237	
AxB1	66	0.008	-0.371	
ABD	85	0.013	-0.509	
A	51	0.005	-0.582	0.298
B1	56	0.006	-0.261	0.385
B2	36	0.006	-0.912	0.509
D	28	0.014	-0.240	0.489
AxB1	50	0.007	-0.646	0.384
ABD	55	0.013	-0.849	0.515

π, Kimura 2-parameter

Table S4. Properties of Sequences Types (STs) and ST complexes

Grouping	Number		Common Properties
	Isolates	STs	
ST3 Cplx	14	4	EHEC/EPEC; HUS/diarrhoea; O103:H2/O111/O128
ST6 Cplx	12	4	EIEC; O124:H30/O152
ST10 Cplx	90	44	global; group A; few pathogens (ETEC/EAEC)
ST11	8	1	EHEC; HUS/diarrhoea; O157:H7/H$^-$
ST13 Cplx	9	7	global; occasional EHEC/diarrhoea
ST14 Cplx	5	3	UPEC; UTI
ST21 Cplx	13	6	EHEC/EPEC; HUS/diarrhoea; O26:H11/O111:H$^-$
ST28 Cplx	4	3	EHEC/EPEC; diarrhoea
ST31 Cplx	4	3	EAEC; diarrhoea
ST62	6	1	NBM/UTI/faecal; O7:K1; OMP 3
ST73 Cplx	11	3	UTI; 025:H1/O6:H1
ST77 Cplx	5	2	Global; non-pathogens
ST90 Cplx	10	4	APEC/EHEC; sepsis/diarrhoea/UTI; O78:K80/O8
ST95 Cplx	22	5	APEC; sepsis/UTI/NBM; O1/O2/O18:K1; OMP 6/9
ST149 Cplx	3	3	*Shigella boydii*
ST152 Cplx	3	3	*Shigella sonnei*
ST155 Cplx	19	11	global; occasional diarrhoea
ST184 Cplx	3	3	Ghana; non-pathogens
ST226 Cplx	4	4	global; diarrhoea
ST243 Cplx	23	12	*Shigella dysenteriae, boydii, flexneri*
ST245 Cplx	13	8	*Shigella flexneri*
ST250 Cplx	6	6	*Shigella dysenteriae, boydii*/ EIEC
ST270 Cplx	11	7	EIEC; O28ac/O136
ST272 Cplx	3	3	EIEC; O144
ST280 Cplx	8	7	EIEC; O143/O167

Reference List

1. Boyd, E.F., and Hartl, D.L. (1997). Nonrandom location of IS1 elements in the genomes of natural isolates of *Escherichia coli*. Mol. Biol. Evol. *14*, 725-732.
2. Sawyer, S.A., Dykhuizen, D.E., DUbose, R.F., Green, L., Mutangadura-Mhlanga, T., Wolczyk, D.F., and hartl, D.L. (1987). Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics *115*, 51-63.

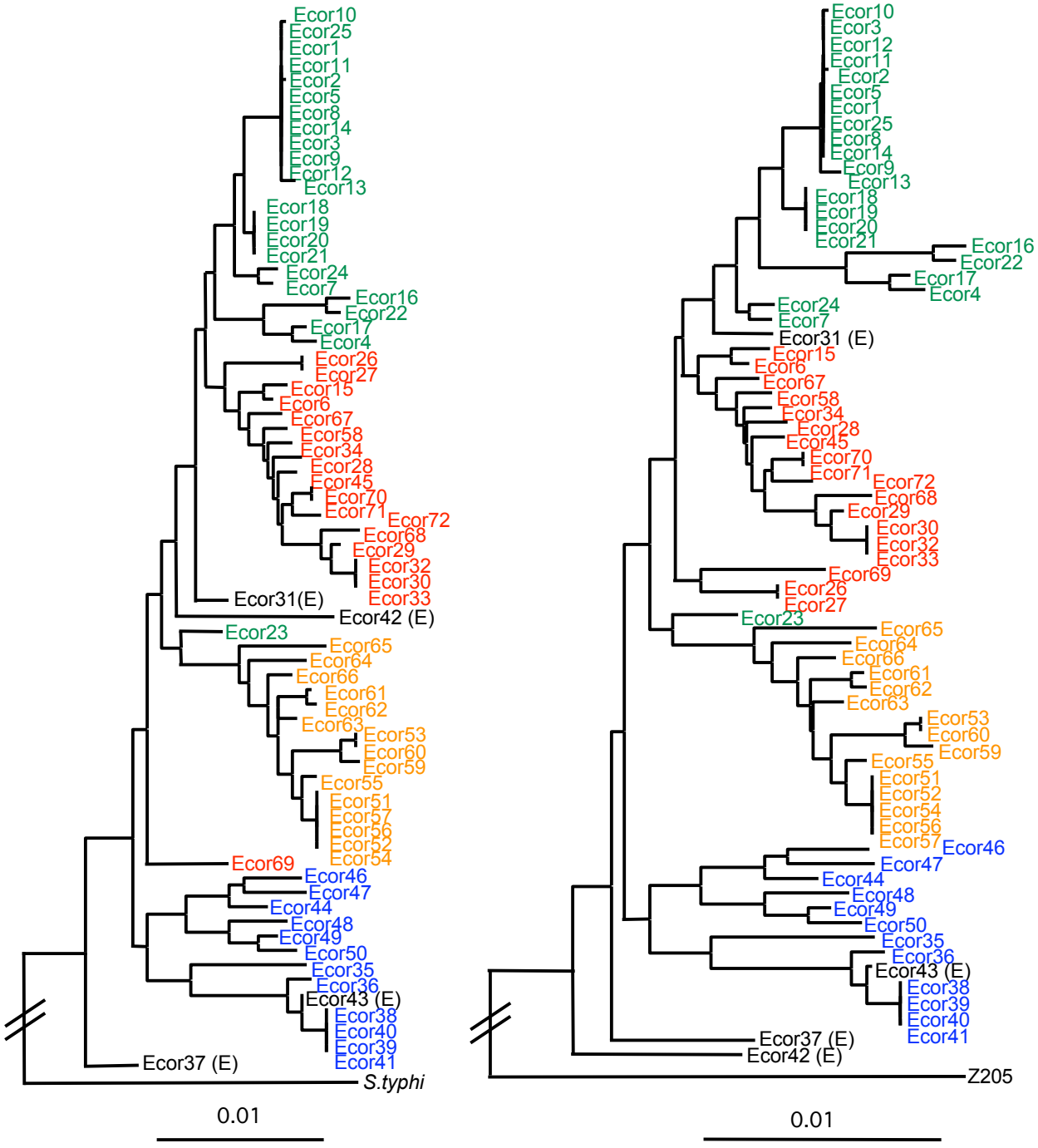

Fig S1. Neighbour-joining trees of the ECOR collection based on the seven concatenated housekeeping genes using the GTR+G+I model. Notice that the trees are rooted with different outgroups. The different strains are colour-coded: group A in green, B1 in red, B2 in yellow, D in blue and E in black.