Erschienen in: FEMS Microbiology Letters ; 350 (2014), 1. - S. 57-64

Phenotype of *htgA* (*mbiA*), a recently evolved orphan gene of *Escherichia coli* and *Shigella*, completely overlapping in antisense to *yaaW*

Lea Fellner[1], Niklas Bechtel[1], Michael A. Witting[2], Svenja Simon[3], Philippe Schmitt-Kopplin[2,4], Daniel Keim[3], Siegfried Scherer[1] & Klaus Neuhaus[1]

[1]Lehrstuhl für Mikrobielle Ökologie, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany; [2]Abteilung Analytische BioGeoChemie, Helmholtz Zentrum München, Neuherberg, Germany; [3]Lehrstuhl für Datenanalyse und Visualisierung, Fachbereich Informatik und Informationswissenschaft, Universität Konstanz, Konstanz, Germany; and [4]Lehrstuhl für Analytische Lebensmittelchemie, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising,Germany

Correspondence: Klaus Neuhaus, Lehrstuhl für Mikrobielle Ökologie, Wissenschaftszentrum Weihenstephan, Technische Universität München, Weihenstephaner Berg 3, D 85350 Freising, Germany.
Tel.: +49 8161 71 3945;
fax: +49 8161 71 4492;
e mail: neuhaus@wzw.tum.de

Abstract

Overlapping embedded genes, such as *htgA/yaaW*, are assumed to be rare in prokaryotes. In *Escherichia coli* O157:H7, *gfp* fusions of both promoter regions revealed activity and transcription start sites could be determined for both genes. Both *htgA* and *yaaW* were inactivated strand specifically by introducing a stop codon. Both mutants exhibited differential phenotypes in biofilm formation and metabolite levels in a nontargeted analysis, suggesting that both are functional despite YaaW but not HtgA could be expressed. While *yaaW* is distributed all over the *Gammaproteobacteria*, an overlapping *htgA* like sequence is restricted to the *Escherichia Klebsiella* clade. Full length *htgA* is only present in *Escherichia* and *Shigella,* and *htgA* showed evidence for purifying selection. Thus, *htgA* is an interesting case of a lineage specific, nonessential and young orphan gene.

Keywords
overlapping genes; *htgA yaaW*; functional characterization.

Introduction

Overlapping embedded genes are considered to be rare in prokaryotes, and only very few have been described (e.g. Silby & Levy, 2008; Tunca *et al.*, 2009; Cheregi *et al.*, 2012). However, the length distribution of overlapping open reading frames in bacteria suggest more of such genes exist (Mir *et al.*, 2012).

The gene *htgA* (high temperature growth, Dean & James, 1991) is located upstream of *dnaK* (James *et al.*, 1993), completely embedded antisense in the hypothetical gene *yaaW* (Fig. 1) and only found in *Escherichia* and *Shigella* (Delaye *et al.*, 2008). Despite its name, a heat shock induction of *htgA* could not be confirmed (Nonaka *et al.*, 2006), and thus, its annotation has been questioned (see Supporting Information, Data S1 for an extended introduction).

We present functional information on both *htgA* and *yaaW,* based on promoter fusions, strand specific single gene knockouts, 5′ RACE and protein expression. Furthermore, the phylogeny of *htgA* is reexamined.

Materials and methods

Transcriptional fusions

Three hundred base pairs (bp) upstream of *htgA* (Z0012), *yaaW* (Z0011) and *yaaI* (Z0013) were PCR amplified (for primers, see Table S1) using *E. coli* O157:H7 EDL933 (EHEC, NC 002655, CIP 106327). The amplicons were cloned upstream *gfp* in pProbe NT (Miller *et al.*, 2000). EHECs with plasmids (verified by sequencing) were grown in LB (Sambrook & Russel 2001) with

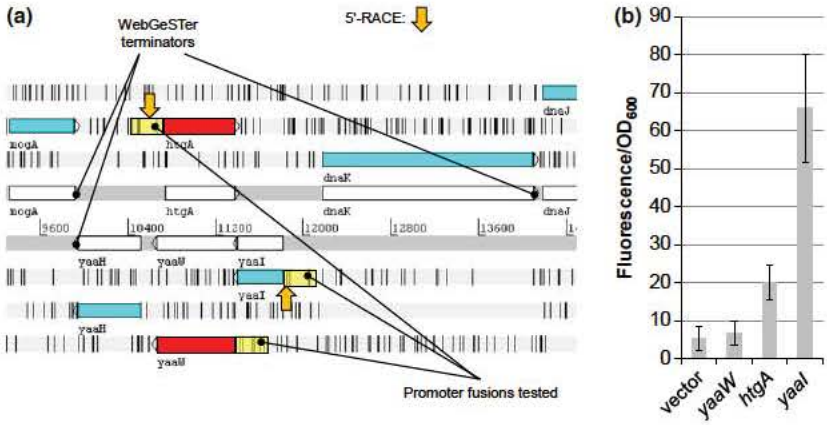

Fig. 1. Overview about the genomic region of *htgA*/*yaaW* and promoter activities. (a) *yaaW* (Z0011) and the overlapping embedded *htgA* (Z0012) are shown on the EHEC EDL933 chromosome. All *yaa* genes are annotated as hypothetical. *dnaK*, downstream of *htgA*, codes for a heat shock protein (chaperone). The black vertical bars indicate stop codons in each reading frame. Yellow boxes of 300 bp length show the promoter regions tested. Orange arrows indicate transcription start sites (+ 1), determined by 5′ RACE. Using the software WEBGESTER, three termination sites could be predicted, as indicated (drawn using Artemis; Rutherford *et al.*, 2000). (b) Fluorescence measured in LB medium for the promoter::*gfp* fusions. The empty *gfp* plasmid serves as the control (vector). The error bars show the standard deviation of three independent measurements with four replicates each.

25 μg mL^{-1} kanamycin. GFP was measured for 1 s of cultures grown in the dark to $OD_{600\ nm}$ = 1, washed once with PBS, and using 200 μL of 1 : 5 and 1 : 10 dilutions (Victor[3], Perkin Elmer). Empty vector control values were measured, and fluorescence was normalized to $OD_{600\ nm}$. The mean of four wells was calculated from three independent experiments.

Transcriptional start sites

5′ RACE was performed using the 5′RACE System for Rapid Amplification of cDNA Ends Version 2.0 (Invitrogen) according to the manufacturer. For *htgA*, the pProbe NT plasmid with an inserted putative promoter region was used, and transformed cells were grown in LB. For *yaaW*, the bacteria were grown in 1 : 10 diluted LB medium at pH6 with 200 mg L^{-1} Na nitrite (R. Landstorfer, S. Simon, S. Schober, D. Keim, S. Scherer & K. Neuhaus, unpublished data) to induce *yaaW*. After gel electrophoresis, the most intense bands were purified (Invisorb® Fragment CleanUp, STRATEC, Berlin), used as template for subsequent amplification and sequenced using nested primers (LGC Genomics, Berlin).

Deletion mutants

For Δ*htgA* and Δ*yaaW*, two DNA fragments were amplified, up and downstream of the site to be mutated, enclosing the mutated site. Both amplicons are used in the subsequent reaction, using the two nonoverlapping primers, to recreate the gene with the mutation. The final product was cloned into pMRS101 (Sarker & Cornelis, 1997). The high copy ori was removed, and the plasmids transferred to *E. coli* CC118λpir (Manoil & Beckwith, 1985). After verification by sequencing, they were transferred to *E. coli* SM10λpir (Miller & Mekalanos, 1988) for mating. EDL933 NalR (spontaneous mutation) and the respective SM10λpir plus the modified pMRS101 were mixed and plated on LB agar (24 h, 30 °C). Cells were resuspended again and plated on LB agar with 30 μg mL^{-1} streptomycin and 20 μg mL^{-1} nalidic acid. Correct plasmid integration after a first cross over was checked by PCR. Second cross over events, resulting in plasmid loss, either restore wild type or create the mutation. Thus, bacteria were grown without selection to $OD_{600\ nm}$ = 0.8 and plated on LB agar without NaCl plus 10% sucrose for *sacB* counter selection. Desired mutants were identified using PCR.

Biofilm assays

Biofilm experiments were conducted according to Domka *et al.* (2007). A culture grown in M9 minimal medium (Sambrook & Russel 2001) was diluted to $OD_{600\ nm}$ = 0.05. Flat bottom wells of a microtiter plate (Greiner Bio One, Germany) were filled with 100 μL and incubated 24 or 48 h without shaking at 30 °C or 37 °C. $OD_{600\ nm}$ was measured (Victor[3]). The planktonic cells were removed, and each well was carefully washed with water. Staining was achieved using 135 μL 0.1% crystal violet (20 min, RT). After washing thrice with water and air drying, the stain was solubilized in 95% ethanol, transferred to a new plate and the absorbance at 600 nm was measured (Victor[3]). The mean was calculated

(10 wells, three biological replicates) after subtracting zero controls (medium only).

Protein expression

Amplicons of *htgA* and *yaaW* were cloned into pBAD/Myc His C (Invitrogen). EHEC with plasmids (sequenced for verification) were grown in LB with 100 μg mL^{-1} ampicil lin and induced with 0.2% arabinose. Proteins were puri fied according to QIAexpress® Ni NTA Fast Start kit under denaturing conditions (Qiagen). For this, the bacteria were sonicated in the provided lysis buffer. For SDS PAGE (15%), Laemmli buffer was added, and the sample dena tured for 5 min at 95 °C. PageRuler Protein Ladder (Fermentas) was used as marker. After electrophoresis, the proteins were electroblotted (20 min, 120 mA) to an activated PVDF membrane (Amersham). Subsequently, the membrane was blocked, incubated with mouse anti human c myc antibodies (BD Biosciences), washed, incu bated with alkaline phosphatase anti mouse chimera anti bodies (Dianova, Hamburg), washed again, equilibrated and incubated in buffer supplemented with BCIP/NBT.

Metabolome assays

Metabolites were profiled using Ion cyclotron resonance Fourier transform Mass spectrometry (ICR FT/MS) on a Bruker solariX with a 12 T magnet (Bruker Daltonics, Bremen). Three biological replicate cultures of wild type, Δ*htgA*, and Δ*yaaW* were grown shaking in 1 : 2 diluted LB to $OD_{600\ nm}$ = 1. Cultures were vacuum filtered using HVLP filters (0.45 μm; Millipore). The bacteria and the filter were flash frozen in liquid nitrogen and extracted with 50% methanol using a FastPrep (MP Biomedicals) with zirconia beads (0.1 mm, and a few beads of 2 mm diameter) three times for 45 s at 6.5 m s^{-1}. Samples were centrifuged, filtered (0.22 μm), diluted 1 : 20 with 70% MeOH, and infused at 120 μL h^{-1}. ICR FT/MS was externally calibrated on clusters of arginine (10 ppm in 70% MeOH). A time domain transient of 2 megawords was used, and 300 scans were accumulated for one spec trum. Spectra were internally calibrated with an error of ≤ 0.1 ppm, exported with a signal to noise ratio of 3, and aligned within a 1 ppm window. Putative metabolites were annotated using MassTRIX (Wägele *et al.*, 2012). Only masses found in all replicates were considered and analyzed in Genedata Expressionist for MS 7.6 (Genedata, Martinsried).

Bioinformatics

Promoters were searched by BPROM (Softberry Inc., New York) and terminators by WEBGESTER DB (Mitra *et al.*, 2011). Microarray data were accessed from the Gene Expression Database (GENEXPDB, http://genexpdb.ou.edu/index.php, see Table 1).

Sequences were searched with BLASTP or TBLASTN (NCBI, http://blast.ncbi.nlm.nih.gov/Blast.cgi, default parameters)

Table 1. Differential expression of *htgA* or *yaaW* detected in microarray experiments

Ratio *htgA*	Ratio *yaaW*	Accession	OUID	Condition	Control
−3.233	−3.675	GSE7	19	Δ*tnaA*, minimal + 50 μg mL^{-1} Trp	Genomic DNA
−3.467	n.v.	GSE533	51	High temperature evolved line 42 1	Ancestor
−2.795	−3.820	GSE7	20	Δ*tnaA*, minimal + 50 μg mL^{-1} Trp	Genomic DNA
−1.717	n.v.	GSE4383	371	Δ*rraA*, $OD_{600\ nm}$ 0.3	WT
−2.715	−0.343	GSE20413	1233	LB medium, Δ*qseD*, K12 strain	WT
−1.506	−1.506	GSE4383	374	*rne* depletion, strain KSL2000	Untreated
3.575	−1.489	GSE10345	875	100 μg bicyclomycin, strain O157:H7	Untreated
2.694	−0.109	GSE 16565	1123	Methyl methane sulfonate (MMS), strain WS3110	Untreated
1.376	2.011	GSE4375	345	Anaerobic, M9 + glucose, $OD_{600\ nm}$ 1.3	Aerobic, $OD_{600\ nm}$ 0.4
2.658	0.868	GSE12831	998	Δ*qseE*	WT
1.169	2.378	GSE7885	750	Δ*rpoS* in exponential phase	WT
2.475	1.205	GSE7439	716	Epinephrine, VS94 strain	Untreated
3.534	0.15	GSE4394	403	6 min postrifampicin, K10 strain	Untreated
n.v.	1.95	GSE13666	1035	Evolved *rpoS*+, strain DMS1735	Ancestor
1.98	2.031	GSE4376	351	Anaerobic, M9 + glucose + fumarate, $OD_{600\ nm}$ 1.3	Aerobic, $OD_{600\ nm}$ 0.4
n.v.	2.056	GSE4569	440	UV irradiated 1 h	Untreated
4.7	−0.442	GSE10345	874	100 μg bicyclomycin, strain MG1655	Untreated
2.62	1.649	GSE3905	207	Biofilm after 15 h of culturing	Biofilm after 4 h
6.043	−0.567	GSE10345	876	100 μg bicyclomycin, strain MDS42	Untreated

Data are from GENEXPDB. Upregulation is shown with positive numbers, downregulation with negative numbers; WT is wild type; n.v., no value given. References are listed under the respective accession number. Note that *yaaW* and *htgA* are treated as synonyms in GENEXPDB, despite differential expression values.

using YaaW (Z0011) as query (Table S2). The evolution ary history of all species was inferred using the software package MEGA5 with a concatemer of 16s rRNA gene, *atpD*, *adk*, *gyrB*, *purA*, and *recA* by Minimum Evolution using p distance. The bootstrap consensus was inferred from 1000 replicates (Tamura *et al.*, 2011). For some strains, not all sequences were available, and thus close relatives were used as surrogate, for example, some genes of *Comamonas testosteroni* CNB 2 were used for the *yaaW* bearing strain ATCC 11996. The presence of *htgA* was detected using pairwise BLASTP alignments with *htgA* (Z0012) as query (starting from the first GTG).

htgA/*yaaW* sequences were examined for their nonsyn onymous over synonymous rate ratio ω as described (Sabath *et al.*, 2008; Sabath & Graur, 2010) including correction for multiple testing according to Benjamini & Hochberg (1995), after omitting alignment gaps (Tamura *et al.*, 2011).

Results and discussion

Transcription of *htgA* and *yaaW*

5′ RACE determined the major 5′ end of the + 1 tran scription start of *htgA* to be 135 bp upstream. However, minor sites might be present, since Missiakas *et al.* (1993) found a site 82 bp upstream; others were predicted 98 (BProm) or 114 bp (Tutukina *et al.*, 2007) upstream of the CTG start codon of *htgA*.

The upstream region of *htgA* was successfully tested for promoter activity using a promoterless *gfp* reporter. No ter minator could be detected directly downstream of *htgA* but was detected downstream of *dnaK* (Fig. 1). Recently, strand specific transcriptome sequencing showed that *htgA* is transcribed, albeit weakly, at some nonlaboratory growth conditions only (R. Landstorfer, S. Simon, S. Schober, D. Keim, S. Scherer & K. Neuhaus, unpublished data).

The 5′ RACE major transcription start site of *yaaW* is 32 bp upstream of *yaaI*, but a minor site, 107 bp upstream of *yaaW*, was also detected. Testing both puta tive promoter regions, we found *yaaI* to show promoter activity, but *yaaW* was similar to the empty vector con trol (Fig. 1b). A terminator was predicted by WEBGESTER downstream of *yaaH* or, in antisense at the same posi tion, downstream of *mog*. This suggests that *yaaW* is most likely organized as operon *yaaIWH* in EHEC and transcribed from the *yaaI* promoter and terminated downstream of *yaaH*.

Interestingly, data from GENEXPDB indicate that *htgA* and *yaaW* are expressed differentially in *E. coli* strains under certain experimental conditions (see Table 1), clearly prohibiting *htgA* synonymizing with *yaaW*, which has been performed in some databases.

Translation of *htgA* and *yaaW*

HtgA and YaaW were expressed in EDL933 using a plas mid that generates concomitant *myc* and His tag fusions. Proteins were prepurified using the his tag and detected on Western blots using the myc tag. YaaW (30 kDa) was detectable, but no band for HtgA was found (Fig. 2), which is in accordance with Narra *et al.* (2008). Thus, the protein might be unstable and difficult to discover. Miss iakas *et al.* (1993) presented a 21 kDa gene product by ^{35}S labeling, which is a more sensitive approach.

Phenotypes of Δ*htgA* and Δ*yaaW* mutants

Previous work always used a double knockout mutant. We created strand specific deletion mutants for the first time, in which only *htgA* or *yaaW* was interrupted (Fig. 3). The annotated *htgA* start codon is CTG, which is quite rare for bacteria. The next GTG is more likely to be the start codon. Counting from there, *htgA* has 525 bp (or 174 amino acids); our *htgA* knock out terminates either product. By introducing a single point mutation to create a stop in one frame, we minimized the disturbance of the other, as the mutations are synonymous in the lat ter (Tunca *et al.*, 2009). For the first time, it was possible to distinguish effects of Δ*htgA* from Δ*yaaW*.

Both mutants showed no difference in their growth compared with wild type at 37 °C or after temperature shift from 30 °C to 45 °C (Fig. 4a). As no heat shock phe notype of Δ*htgA* could be confirmed (as found before, Nonaka *et al.*, 2006), *htgA* should no longer be annotated as heat shock gene. In minimal medium, biofilm forma tion of Δ*htgA* or Δ*yaaW* was reliably increased when incu bated for 48 h at 37 °C (Fig. 4b). This is in accordance

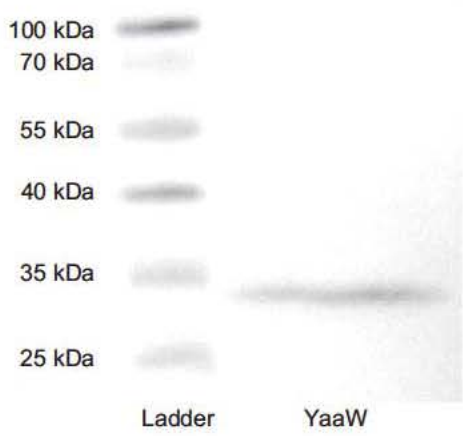

Fig. 2. Western blot detection of YaaW protein. *yaaW* was fused to a myc his tag and overexpressed in EHEC. The fusion protein was purified with Ni NTA and detected on the Western Blot using a myc antibody.

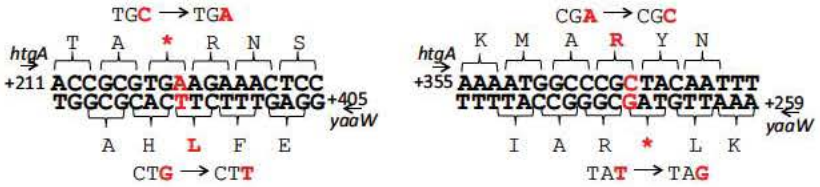

Fig. 3. Point mutations to interrupt *htgA* and *yaaW*. The mutated bases are shown in red. Note that this technique destroys only one open reading frame at a time, but leaves the other one intact using a synonymous codon.

with Domka *et al.* (2007), who found a threefold increase in biofilm formation for *E. coli* K12 in a *htgA*/*yaaW* double mutant. We speculate that the higher increase compared with our experiments might be due to additive effects of both genes in the double mutant compared with each single one. We therefore suggest to rename *htgA* to *mbiA* (modifier of biofilm).

Metabotypes of Δ*htgA* and Δ*yaaW* mutants

As no difference in growth could be found, we measured the metabotypes. Metabolite changes could still be detect able even though they may not manifest in growth (Raamsdonk *et al.*, 2001). Δ*htgA*, Δ*yaaW*, and wild type were subjected to nontargeted metabolomics using ICR FT/MS. Indeed, twenty two different metabolites (puta tively annotated, see Table S3) between the strains were found significantly changed ($P \leq 0.01$). When comparing Δ*htgA* to wild type, we found four differences, comparing Δ*yaaW* to wild type, 14 differences, and comparing Δ*htgA* to Δ*yaaW*, four differences. In both mutants, all metabo lites were decreased compared with wild type. The differen tial changes provide evidence that both reading frames are functional. The majority of changes were associated with fatty or amino acid metabolism. Neither *htgA* nor *yaaW* appear to be directly involved in the cellular metabolism and any functional explanation is as yet highly speculative.

Is *htgA* an RNA only?

Instead of being protein coding, *htgA* could produce a regulatory (metabolite binding) or antisense RNA. This is considered unlikely as several metabolites are affected. More importantly, antisense RNA regulation is achieved by base pairing of longer stretches between the antisense and target RNA (Lasa *et al.*, 2012), but we engineered single base substitutions, which should not cause any detectable differences in pairing.

Taxonomical distribution and evolution of *htgA*

yaaW homologs are present in a variety of bacteria (Fig. 5, Table S2), but a complete *htgA* frame is present only in *Escherichia* and *Shigella*. A minority of *Salmonella* contains *yaaW*, but *htgA* is always a pseudogene in those species and interestingly in each case disrupted at the same positions.

Evolution of *yaaW* is restricted when it contains an overlapping *htgA* frame (Delaye *et al.*, 2008). The rate

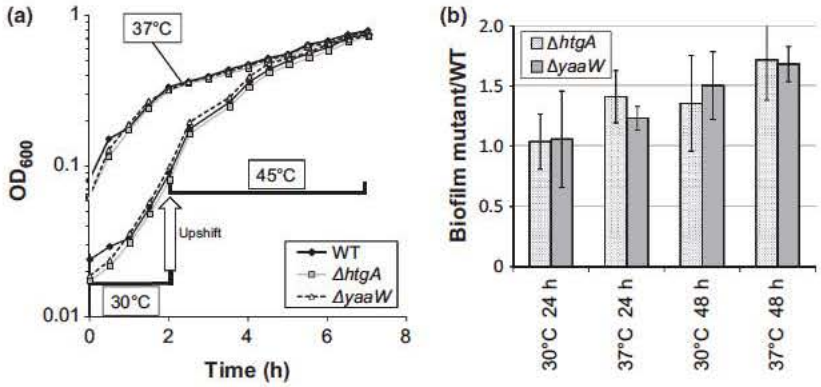

Fig. 4. Phenotypes of wild type, Δ*htgA*, and Δ*yaaW*. (a) Growth curves recorded at different temperatures. The upper curves show the growth at 37 °C. For the lower curves, the bacteria were grown first at 30 °C to $OD_{600\ nm}$ 0.1 and subsequently shifted to 45 °C as indicated by the arrow. (b) Biofilm formation of the mutants Δ*htgA* or Δ*yaaW* compared with the wild type grown in M9 minimal medium. Temperature and time regimens are indicated. Both mutants show an increase in biofilm formation compared with the wild type (WT), especially after 48 h at 37 °C. The error bars show the standard deviation of three independent measurements with ten replicates each.

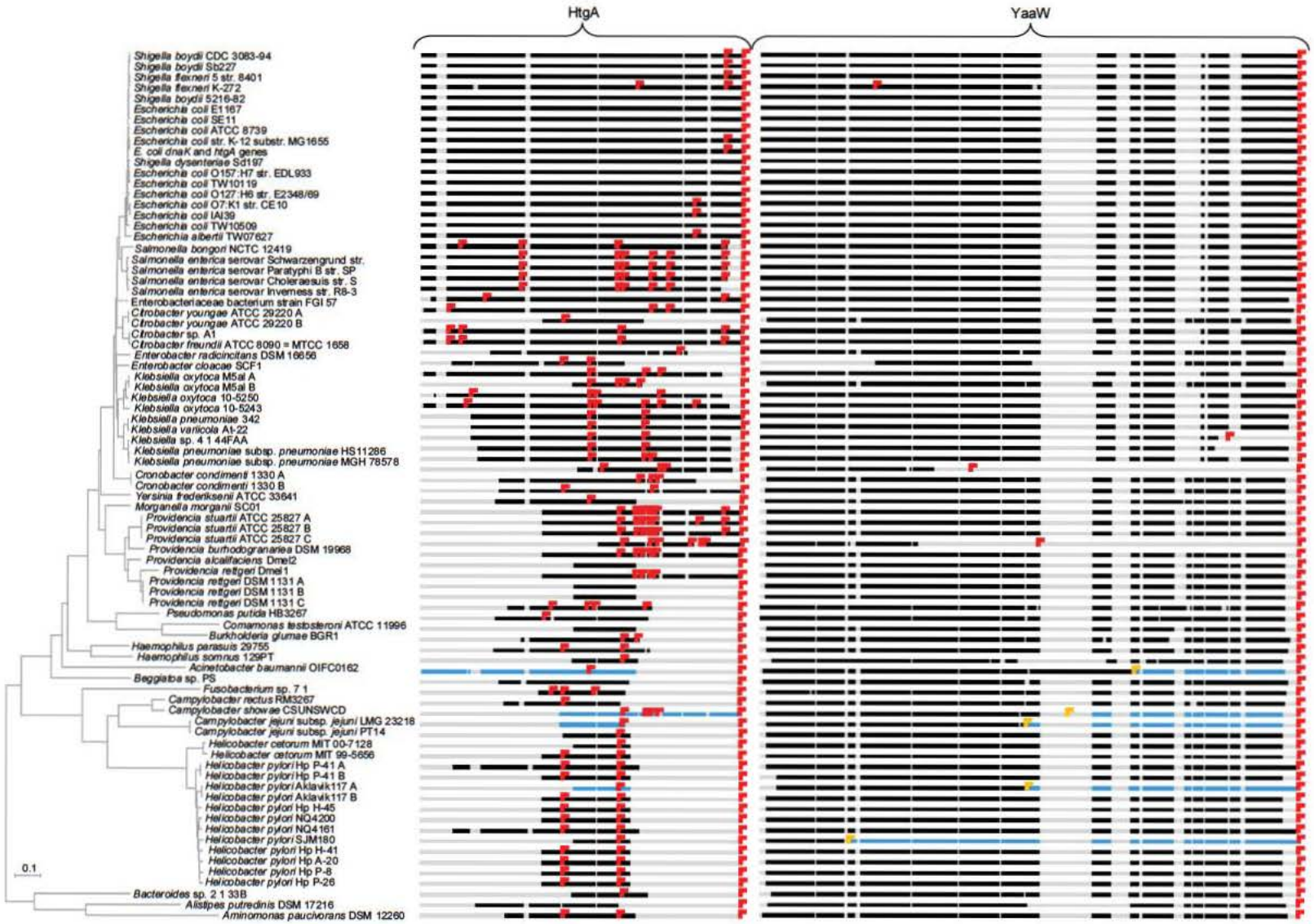

Fig. 5. Phylogenetic distribution of YaaW and HtgA. A representative subset of organisms carrying *htgA/yaaW* gene locus is shown. Left: Phylogenetic tree constructed using the minimum evolution tree method. Middle and right: Graphic representation of HtgA and YaaW. Black, protein sequence which could be aligned to the query (HtgA or YaaW) using BLASTP; red, stop codons; yellow, site of frame shift; blue, shifted sequence after frame shift; gray, sequence could not be aligned using BLASTP with standard settings. Note that some strains contain paralogs of *yaaW*, which are indicated with A, B, or C after the strain name.

between synonymous and nonsynonymous mutations in a gene is used to infer selection. However, embedded genes influence each other, invalidating models used for nonoverlapping genes. Sabath *et al.* (2008) designed a model to estimate the nonsynonymous over synonymous substitution rate of overlapping genes to infer selection, comparing two scenarios: The first makes no assump tions on any selection intensity, the second assumes 'no selection' for the overlap, here *htgA*. In strains in which *htgA* was interrupted, indeed no selection was found. However, the estimation of selection intensities is lim ited in case of low sequence diversity, which is the case for *yaaW* (max. 2.6% on amino acid level). *htgA* is encoded in frame 2 in relation to *yaaW*, which provides the least flexibility for amino acid changes of both (Rogozin *et al.*, 2002). This may partly explain the com paratively low degree of divergence. Despite these limita tions, *htgA* is expected to be under (purifying) selection, and hence functional, in at least 24 strains of *Escherichia* and *Shigella* (Table S4).

We suggest that *htgA* is a young orphan (taxonomically restricted gene), as full length *htgA* is restricted to *Escheri chia* and *Shigella,* originating probably before *Citrobacter* or *Klebsiella* have separated. Orphans seem to be responsible for lineage specific adaptations and most of these are assumed to be evolutionary 'young' genes, showing higher divergence rates, lower expression rates and encode shorter proteins compared to older genes (Tautz & Domazet Loso, 2011). Despite that such genes most likely have no essential function and, therefore, may be prone to be lost again (e.g. in *Salmonella*), *htgA* should be added once again to the genome annotation of *E. coli* as an interesting case of an overlapping gene which emerged recently.

Acknowledgements

This study was funded by the DFG (SCHE316/3 1, KE740/13 1). We would like to thank Luke Tyler for assisting with the language. The authors declare that they have no conflict of interests.

References

Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57**: 289 300.

Cheregi O, Vermaas W & Funk C (2012) The search for new chlorophyll binding proteins in the cyanobacterium *Synechocystis* sp. PCC 6803. *J Biotechnol* **162**: 124 133.

Dean DO & James R (1991) Identification of a gene, closely linked to *dnaK*, which is required for high temperature growth of *Escherichia coli*. *J Gen Microbiol* **137**: 1271 1277.

Delaye L, Deluna A, Lazcano A & Becerra A (2008) The origin of a novel gene through overprinting in *Escherichia coli*. *BMC Evol Biol* **8**: 31.

Domka J, Lee J, Bansal T & Wood TK (2007) Temporal gene expression in *Escherichia coli* K 12 biofilms. *Environ Microbiol* **9**: 332 346.

James R, Dean DO & Debbage J (1993) Five open reading frames upstream of the *dnaK* gene of *E. coli*. *DNA Seq* **3**: 327 332.

Lasa I, Toledo Arana A & Gingeras TR (2012) An effort to make sense of antisense transcription in bacteria. *RNA Biol* **9**: 1039 1044.

Manoil C & Beckwith J (1985) TnphoA: a transposon probe for protein export signals. *P Natl Acad Sci USA* **82**: 8129 8133.

Miller VL & Mekalanos JJ (1988) A novel suicide vector and its use in construction of insertion mutations: osmoregulation of outer membrane proteins and virulence determinants in *Vibrio cholerae* requires *toxR*. *J Bacteriol* **170**: 2575 2583.

Miller WG, Leveau JH & Lindow SE (2000) Improved *gfp* and *inaZ* broad host range promoter probe vectors. *Mol Plant Microbe Interact* **13**: 1243 1250.

Mir K, Neuhaus K, Scherer S, Bossert M & Schober S (2012) Predicting statistical properties of open reading frames in bacterial genomes. *PLoS ONE* **7**: e45103.

Missiakas D, Georgopoulos C & Raina S (1993) The *Escherichia coli* heat shock gene *htpY*: mutational analysis, cloning, sequencing, and transcriptional regulation. *J Bacteriol* **175**: 2613 2624.

Mitra A, Kesarwani AK, Pal D & Nagaraja V (2011) WebGeSTer DB a transcription terminator database. *Nucleic Acids Res* **39**: D129 D135.

Narra HP, Cordes MH & Ochman H (2008) Structural features and the persistence of acquired proteins. *Proteomics* **8**: 4772 4781.

Nonaka G, Blankschien M, Herman C, Gross CA & Rhodius VA (2006) Regulon and promoter analysis of the *E. coli* heat shock factor, σ32, reveals a multifaceted cellular response to heat stress. *Genes Dev* **20**: 1776 1789.

Raamsdonk LM, Teusink B, Broadhurst D *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol* **19**: 45 50.

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL & Koonin EV (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet* **18**: 228 232.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA & Barrell B (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944 945.

Sabath N & Graur D (2010) Detection of functional overlapping genes: simulation and case studies. *J Mol Evol* **71**: 308 316.

Sabath N, Landan G & Graur D (2008) A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS ONE* **3**: e3996.

Sambrook J & Russel DW (2001) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York, NY.
Sarker MR & Cornelis GR (1997) An improved version of suicide vector pKNG101 for gene replacement in Gram negative bacteria. *Mol Microbiol* **23**: 410 411.
Silby MW & Levy SB (2008) Overlapping protein encoding genes in *Pseudomonas fluorescens* Pf0 1. *PLoS Genet* **4**: e1000094.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731 2739.
Tautz D & Domazet Loso T (2011) The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692 702.
Tunca S, Barreiro C, Coque JJ & Martin JF (2009) Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). *FEBS J* **276**: 4814 4827.
Tutukina MN, Shavkunov KS, Masulis IS & Ozoline ON (2007) Intragenic promotor like sites in the genome of *Escherichia coli* discovery and functional implication. *J Bioinform Comput Biol* **5**: 549 560.
Wägele B, Witting M, Schmitt Kopplin P & Suhre K (2012) MassTRIX reloaded: combined analysis and visualization of transcriptome and metabolome data. *PLoS ONE* **7**: e39860.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Extended introduction.
Table S1. Primer used in this study.
Table S2. Bacterial species with *yaaW* and sequences used.
Table S3. MassTRIX annotation of significantly changed mass fingerprints.
Table S4. Results of non synonymous over synonymous rate ratios ω.