REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	New Reprint	-

4. TITLE AND SUBTITLE
Characterization and evolution of Salmonella CRISPR-Cas systems

5a. CONTRACT NUMBER
W911NF-11-1-0442

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
N. Shariat, R. E. Timme, J. B. Pettengill, R. Barrangou, E. G. Dudley

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Pennsylvania State University
Office of Sponsored Programs
110 Technology Center Building
University Park, PA 16802 -7000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60012-LS.5

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Prokaryotic CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR-associated genes) systems provide adaptive immunity from invasive genetic elements and encompass three essential features: (i) cas genes, (ii) a CRISPR array composed of spacers and direct repeats and (iii) an AT-rich leader sequence upstream of the array. We performed in- depth sequence analysis of the CRISPR-Cas systems in .600 Salmonella, representing four clinically prevalent serovars. Each CRISPR-Cas feature is extremely conserved in the Salmonella, and the CRISPR1 locus is more highly conserved than CRISPR2. Array composition is serovar- specific, although no

15. SUBJECT TERMS
CRISPR, Salmonella, evolution

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Edward Dudley
UU	UU	UU	UU		19b. TELEPHONE NUMBER 814-867-0439

Report Title

Characterization and evolution of Salmonella CRISPR-Cas systems

ABSTRACT

Prokaryotic CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR-associated genes) systems provide adaptive immunity from invasive genetic elements and encompass three essential features: (i) cas genes, (ii) a CRISPR array composed of spacers and direct repeats and (iii) an AT-rich leader sequence upstream of the array. We performed in- depth sequence analysis of the CRISPR-Cas systems in .600 Salmonella, representing four clinically prevalent serovars. Each CRISPR-Cas feature is extremely conserved in the Salmonella, and the CRISPR1 locus is more highly conserved than CRISPR2. Array composition is serovar- specific, although no convincing evidence of recent spacer acquisition against exogenous nucleic acids exists. Only 12% of spacers match phage and plasmid sequences and self-targeting spacers are associated with direct repeat variants. High nucleotide identity (.99.9%) exists across the cas operon among isolates of a single serovar and in some cases this conservation extends across divergent serovars. These observations reflect historical CRISPR-Cas immune activity, showing that this locus has ceased undergoing adaptive events. Intriguingly, the high level of conservation across divergent serovars shows that the genetic integrity of these inactive loci is maintained over time, contrasting with the canonical view that inactive CRISPR loci degenerate over time. This thorough characterization of Salmonella CRISPR-Cas systems presents new insights into Salmonella CRISPR evolution, particularly with respect to cas gene conservation, leader sequences, organization of direct repeats and protospacer matches. Collectively, our data suggest that Salmonella CRISPR-Cas systems are no longer immunogenic; rather, their impressive conservation indicates they may have an alternative function in Salmonella.

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 60012.5-LS
Characterization and evolution of Salmonella CF...

Block 13: Supplementary Note

Approved for public release; distribution is unlimited.

Microbiology (2015), **161**, 374–386 DOI 10.1099/mic.0.000005

Characterization and evolution of *Salmonella* CRISPR-Cas systems

Nikki Shariat,[1] Ruth E. Timme,[2] James B. Pettengill,[2] Rodolphe Barrangou[3] and Edward G. Dudley[1]

Correspondence
Edward G. Dudley
egd100@psu.edu

[1]Department of Food Science, Center for Molecular Immunology and Infectious Disease and Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA 16802, USA

[2]Center for Food Safety and Applied Nutrition, US Food and Drug Administration, College Park, MD, USA

[3]Department of Food, Bioprocessing and Nutrition Services, North Carolina State University, NC, USA

Prokaryotic CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR-associated genes) systems provide adaptive immunity from invasive genetic elements and encompass three essential features: (i) *cas* genes, (ii) a CRISPR array composed of spacers and direct repeats and (iii) an AT-rich leader sequence upstream of the array. We performed in-depth sequence analysis of the CRISPR-Cas systems in >600 *Salmonella*, representing four clinically prevalent serovars. Each CRISPR-Cas feature is extremely conserved in the *Salmonella,* and the CRISPR1 locus is more highly conserved than CRISPR2. Array composition is serovar-specific, although no convincing evidence of recent spacer acquisition against exogenous nucleic acids exists. Only 12 % of spacers match phage and plasmid sequences and self-targeting spacers are associated with direct repeat variants. High nucleotide identity (>99.9 %) exists across the *cas* operon among isolates of a single serovar and in some cases this conservation extends across divergent serovars. These observations reflect historical CRISPR-Cas immune activity, showing that this locus has ceased undergoing adaptive events. Intriguingly, the high level of conservation across divergent serovars shows that the genetic integrity of these inactive loci is maintained over time, contrasting with the canonical view that inactive CRISPR loci degenerate over time. This thorough characterization of *Salmonella* CRISPR-Cas systems presents new insights into *Salmonella* CRISPR evolution, particularly with respect to *cas* gene conservation, leader sequences, organization of direct repeats and protospacer matches. Collectively, our data suggest that *Salmonella* CRISPR-Cas systems are no longer immunogenic; rather, their impressive conservation indicates they may have an alternative function in *Salmonella*.

Received 17 September 2014
Accepted 12 November 2014

INTRODUCTION

Salmonella enterica is an enteric pathogen and the primary cause of bacterial foodborne illness in the United States (Scallan *et al.*, 2011). It is a tremendously diverse species comprising six subspecies and over 2500 serovars (Grimont & Weill, 2007). *S. enterica* subsp. *enterica* accounts for the majority of clinical cases of salmonellosis and the majority of serovar diversity (~1500 serovars). Serovars (ser.) Enteritidis, Typhimurium, Heidelberg and Newport are collectively responsible for 44 % of illness cases annually (Centers for Disease Control & Prevention, 2011).

Clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated (Cas) systems are found in ~45 % of bacterial genomes (Grissa *et al.*, 2007), including *Salmonella*. Canonically, CRISPR-Cas systems provide an adaptive immune response to bacteriophages and plasmids (reviewed by Bhaya *et al.*, 2011; Wiedenheft *et al.*, 2012). They comprise three major features: a set of *cas* genes, a leader sequence and a CRISPR array (Fig. 1). The CRISPR array, or spacer array, is composed of direct repeat sequences that are interspaced with unique spacer sequences that are typically derived from mobile genetic elements such as bacteriophages and plasmids (Barrangou *et al.*, 2007; Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005). The AT-rich leader sequence lies directly upstream of each

Abbreviations: Cas, CRISPR-associated; CRISPR, clustered regularly interspaced short palindromic repeats; crRNA, CRISPR RNA; DRV, direct repeat variant; LCA, last common ancestor.

Three supplementary tables and five supplementary figures are available with the online Supplementary Material.

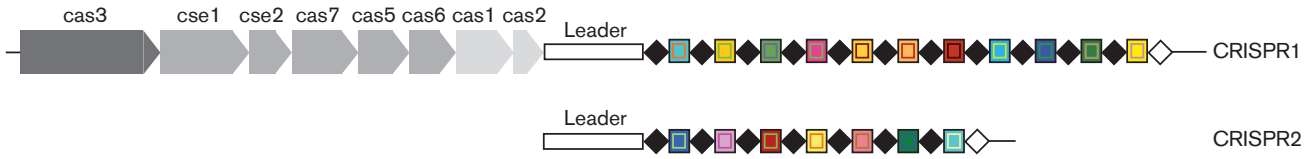

Fig. 1. *Salmonella* CRISPR-Cas loci. *Salmonella* have two CRISPR loci, CRISPR1 and CRISPR2, both encoded on the minus strand. There are eight *cas* genes which are located upstream of CRISPR1, shown as grey boxed arrows. The type I system signature gene, *cas3*, is shown (dark grey). The *cas1* and *cas2* genes are universal, present in all CRISPR-Cas systems (light grey). The remaining *cas* genes are type I-E-dependent. AT-rich leader sequences are situated directly upstream of both CRISPR spacer arrays (open boxes). The spacer array comprises direct repeats (filled diamonds) that are separated by unique spacers (coloured squares). The terminal direct repeats have divergent sequences (open diamond).

array and is thought to function as a promoter (Jansen *et al.*, 2002; Pul *et al.*, 2010). The spacer array is transcribed and processed into small CRISPR RNAs (crRNAs) each of which consists of the spacer flanked by portions of the direct repeat (Brouns *et al.*, 2008; Hale *et al.*, 2008, 2009; Lillestøl *et al.*, 2006). In concert with some Cas proteins, the mature crRNA is targeted to complementary nucleic acids, such as an invading phage genome, resulting in target DNA degradation (Garneau *et al.*, 2010).

CRISPR-Cas systems adapt by acquiring new spacers at the leader proximal end (Barrangou *et al.*, 2007), providing polarity to the array with older spacers residing at the leader distal end and newer spacers closest to the leader sequences (Horvath *et al.*, 2008; Pourcel *et al.*, 2005). The cognate spacer sequences in the target nucleic acid are known as protospacers (Deveau *et al.*, 2008). Hallmarks of an adaptive immune locus include conserved *cas* genes and leader sequences, plus CRISPR arrays that are divergent between distinct but closely related strains, due to recent spacer acquisition.

Salmonella has two CRISPR loci, CRISPR1 and CRISPR2 (Fig. 1), separated by ~16 kb and which share the same consensus direct repeat sequence (29 nt); the spacers are 32 nt in length. It is well established in *Salmonella* that the overwhelming majority of CRISPR allelic polymorphisms within a serovar arise from deletion or duplication of direct repeat–spacer units, rather than acquisition of new spacers (Fabre *et al.*, 2012; Liu *et al.*, 2011a, b; Shariat *et al.*, 2013a). There are eight *cas* genes, *cas3*, *cse1*, *cse2*, *cas7*, *cas5*, *cas6e*, *cas1* and *cas2*, which are characteristic of a type I-E CRISPR-Cas system (Makarova *et al.*, 2011).

To date, the CRISPR loci from several hundred isolates of *Salmonella* have been analysed with the aim of developing subtyping protocols (DiMarzio *et al.*, 2013; Fabre *et al.*, 2012; Liu *et al.*, 2011a; Shariat *et al.*, 2013a, b) or gaining a better understanding of *Salmonella* phylogeny (Fricke *et al.*, 2011; Pettengill *et al.*, 2014; Timme *et al.*, 2013). Using whole genome assemblies, Pettengill *et al.* (2014) provided a bird's eye view of CRISPR-Cas biology in *Salmonella* across 64 serovars, showing two distinct *cas* gene profiles and a high diversity in length of both CRISPR arrays between different serovars.

By using sequence analysis of several distinct isolates of four clinically relevant serovars, Enteritidis, Typhimurium, Newport and Heidelberg, our goal here was to gain a deeper evolutionary understanding of all components of the *Salmonella* CRISPR-Cas system. Our data show that both the *cas* operon and leaders are well conserved in all serovars, as are the arrays, with respect to spacer content and organization. We observe a lack of spacer acquisition and this, plus the low number of protospacers identified in bacteriophage and plasmid sequences, suggests that these elements do not provide an immune function in *Salmonella*.

METHODS

Bacterial isolates and sequence analysis. We analysed the CRISPR1 and CRISPR2 arrays from 400 clinical *Salmonella* isolates from our collection that included 141 ser. Enteritidis, 84 Typhimurium, 86 Newport and 89 Heidelberg (Shariat *et al.*, 2013a, b, c). These isolates were collected over 5 years and generally one isolate per serovar per month was analysed. In our previous work, CRISPR sequences were combined with multi-locus virulence sequence typing as a molecular subtyping application but the CRISPR arrays were not analysed in depth. Here, the spacers were visualized using a macro, as previously reported (Liu *et al.*, 2011a). The accession numbers for CRISPR alleles are listed in Table S1 (available in the online Supplementary Material). For simplicity, we refer to individual CRISPR alleles as arrays; i.e. allele 66 from our previous publications is referred to as array 66.

In total, 206 isolates [97 ser. Enteritidis; 45 ser. Heidelberg; 53 ser. Newport (21 ser. Newport Lineage II and 31 Lineage III) and 12 ser. Typhimurium (including three ser. Typhimurium monophasic variants, i, 4,[5], 12:i:−)] were sequenced as part of a US Food and Drug Administration initiative (Pettengill *et al.*, 2014; Timme *et al.*, 2013). Accession numbers for the whole genome sequences are listed in Table S2. The *cas* genes and leader sequences were extracted from these assemblies. These are draft genomes; where we were unable to determine the full sequence of one or more *cas* genes (due to contig gaps or presence/absence of a base within a homopolymeric region), we removed the entire isolate from analysis. The leader sequence of CRISPR1 was defined as the sequence between *cas2* and the first direct repeat in the array (96 bp). The CRISPR2 leader was subsequently defined as the 96 bp sequence upstream of the first direct repeat. We located the two leaders in a single ser. Typhimurium isolate and then used the program BLAST (Altschul *et al.*, 1990) to subsequently identify others in our dataset. Given the lack of sequence similarity between ser. Newport-II CRISPR1 leader sequences and the other serovars, we manually curated the leader sequences by extracting the sequence bound by *cas2* and the first direct repeat.

All sequence analyses and alignments were done using the DNA Star Lasergene 11 suite (DNA Star). The nucleotide identity of the *cas* operon was defined as the percentage of identical nucleotides across the whole operon within a serovar. Similarly, the nucleotide identity of *cas* genes was defined as the percentage of identical nucleotides occurring in a particular *cas* gene across all 208 isolates.

Determining spacer matches/identifying protospacers. Putative protospacer matches were identified using CRISPRTarget (Biswas *et al.*, 2013). We considered matches to be $\geqslant$84 % (minimum of 27/32 matching nucleotides). To determine whether protospacers that were annotated as genomic were in fact within prophage regions, we extracted the sequence 20 kb upstream and 20 kb downstream of the protospacer and analysed this sequence using the program PHAST (http://phast.wishartlab.com/) (Zhou *et al.*, 2011).

RESULTS

Diversity and overview of *Salmonella* CRISPR arrays

Salmonella have two CRISPR loci, CRISPR1 and CRISPR2 (Fig. 1). The CRISPR spacer array data were derived from Sanger sequencing of CRISPR spacer arrays that were PCR-amplified (Shariat *et al.*, 2013a, b, c). All the CRISPR1 and CRISPR2 arrays identified are shown in Figs 2 and 3, with the direct repeat sequences removed for clarity. We found 61 and 68 different arrays for CRISPR1 and CRISPR2, respectively, among the four serovars (ser. Enteritidis, Newport, Heidelberg and Typhimurium; Table 1). Serovar Typhimurium had the largest number of different arrays for both loci. For each serovar, the most frequent array found at each locus is indicated in Figs 2 and 3.

In total among all arrays, we identified 179 unique spacers. The mean number of unique spacers in an array was 16 (CRISPR1) and 20 (CRISPR2). The smallest array seen in a single isolate contained two spacers and three direct repeats (ser. Typhimurium, CRISPR1 array 131). Interestingly, these two spacers represent the oldest and newest spacers (Fig. 2). The largest CRISPR arrays contained 34 unique spacers and 35 direct repeats (four ser. Typhimurium CRISPR2 arrays: 164, 173, 179 and 207; Fig. 3). On average, ser. Enteritidis has the smallest and also the fewest number of different CRISPR arrays (Table 1).

Analysis of CRISPR array differences

Spacer loss. The majority of CRISPR1 array differences (54/61 arrays) occur due to loss of one or more spacers, for example in ser. Enteritidis CRISPR1 arrays 2, 15 and 69 (Fig. 2). Although spacer loss also occurs in most CRISPR2 arrays (49/68), other genetic alterations also occur that define array differences (see below). Serovar Heidelberg is the only serovar in which all CRISPR2 array disparities are due to loss of internal spacers. Spacer loss more commonly involves loss of two or more contiguous spacers, rather than a single spacer (Figs 2 and 3).

To determine any bias toward spacers being lost from the leader proximal versus distal ends of the array, we calculated the spacer loss events and performed a *t*-test. Loss of contiguous spacers was considered a single event. We found no significant difference between spacer loss in one-half of the array versus the other half ($P>0.1$).

Spacer duplication. Duplication of spacers was only observed in CRISPR2 and in all serovars except ser. Heidelberg. Spacer duplication occurs as a single copied unit [such as ser. Newport-II CRISPR2 spacer (sp) 22] or a single spacer duplicated multiple times (e.g. ser. Enteritidis CRISPR2 sp9 and ser. Typhimurium CRISPR2 sp26). In ser. Typhimurium (arrays 181 and 205) there is a region of duplication involving seven spacers (sp6 and 7 and sp8, 9–13) that presumably encompasses two independent duplication events.

SNPs. There are only three cases of SNPs occurring in a spacer: ser. Enteritidis CRISPR1 sp2 (this spacer is found with one or two SNPs as indicated in Fig. 2), ser. Typhimurium CRISPR2 sp12 and ser. Newport-III CRISPR2 sp9. With the exception of the last named, these SNPs are seen in multiple isolates (Figs 2 and 3). We found several SNPs within the direct repeats (see below) although only two of these were not conserved (ser. Enteritidis CRISPR1 array 66 and ser. Newport-III CRISPR2 array 145).

Unique spacers. The final demonstration of array differences is the presence of unique spacers that only exist in one strain. We found six unique spacers within our isolate collection: two were positioned at the leader proximal end of the array and four were found internally. Three of the unique spacers were in ser. Typhimurium CRISPR1 loci; array 143 contains both a unique spacer at the leader proximal position as well as an internal unique spacer (sp28) and array 134 also has a leader proximal unique spacer (Fig. 2). Unique leader proximal spacers may be considered putative examples of spacer acquisition. Serovar Newport-II CRISPR2 allele 137 contains two unique spacers (sp15 and 16) that are positioned internally and not found in other isolates (Fig. 3).

Similarities between ser. Typhimurium, Heidelberg and Newport

Serovars Typhimurium and Heidelberg have very similar CRISPR loci (Figs 2 and 3); 76 % (CRISPR1) and 100 % (CRISPR2) of the spacers from ser. Heidelberg arrays are found in ser. Typhimurium and their order within the arrays is identical. The unique spacers in CRISPR1 and the unique ser. Typhimurium spacers in CRISPR2 are seen at the leader proximal end of the array, consistent with what is understood about CRISPR adaptation and evolution (Barrangou *et al.*, 2007). Considering the extensive overlap of CRISPR2 spacers, it is somewhat surprising that no identical CRISPR2 arrays are shared between these two serovars. Additionally, 35 % of CRISPR2 spacers from ser. Newport-III are also found in ser. Typhimurium.

The anchor spacer (sp1) is the furthest from the leader and is the oldest spacer in terms of acquisition. This spacer is

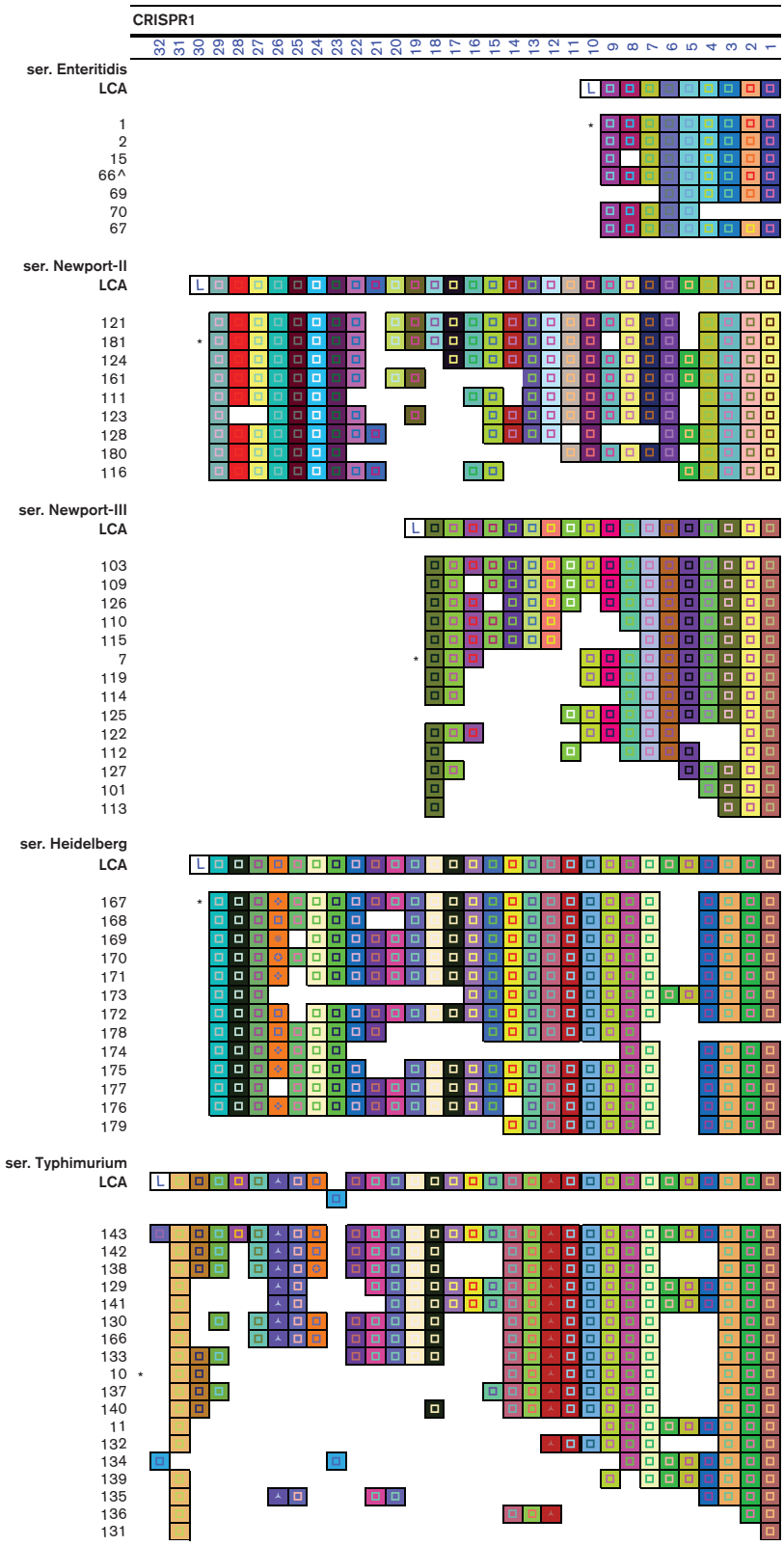

Fig. 2. Spacer organization in CRISPR1. Graphic representation of the unique CRISPR1 arrays from 400 *Salmonella* isolates analysed. For clarity, the direct repeat sequences have been removed and only the spacer sequences are represented. The direction of the spacers is shown 5′–3′, with respect to the leader; the leader is represented by a boxed 'L'. Each unique spacer is represented by a unique combination of background colour and the colour and shape of the object in the foreground. The spacers are aligned and the gaps represent the absence of a particular spacer. The putative LCA for each serovar is shown on the first line of each serovar group. Unique arrays are given a numeric identifier, which is listed to the left of the respective CRISPR array. The array that occurs most frequently for each serovar is shown with an asterisk directly to the left of the array. This was the only case of an SNP occurring in a direct repeat that defined a single array. The bold line upstream of sp10 in ser. Heidelberg and Typhimurium represents a truncated direct repeat between sp10 and sp11. SNPs in ser. Enteritidis (sp2) are shown by variations in colour of the box, and the presence of repeated elements in ser. Typhimurium and Heidelberg (sp24 and sp26, respectively) is shown by altered foreground shapes.

well maintained within each serovar and is only missing in three CRISPR arrays. In CRISPR1, the anchor spacer is shared between ser. Typhimurium, Heidelberg and Newport-III, although Newport-III contains an SNP. In CRISPR2, a conserved anchor group of the three oldest spacers is also shared between these three serovars.

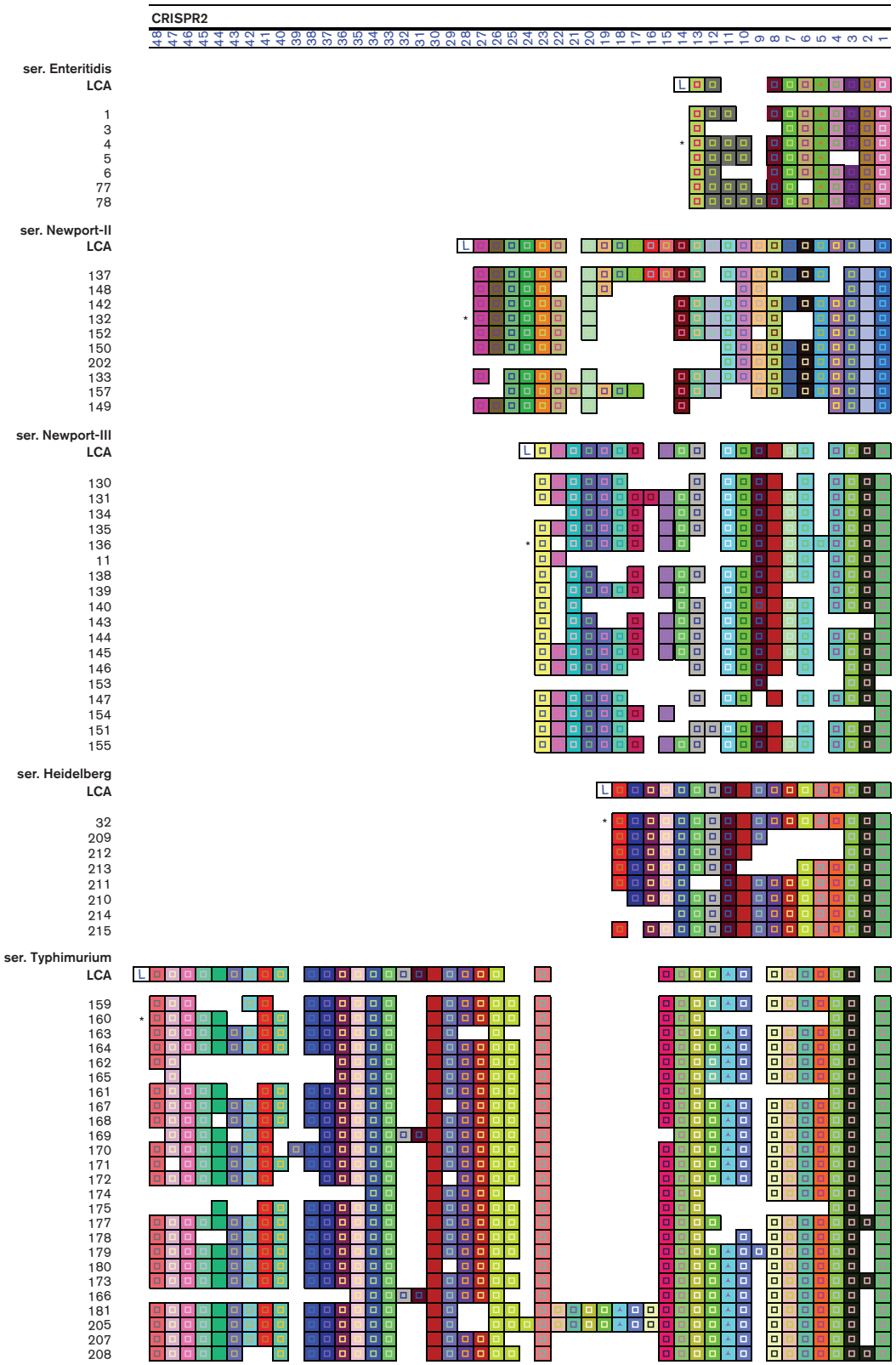

Fig. 3. Spacer organization in CRISPR2. The data are presented as in Fig. 2. In some ser. Newport-III isolates, sp15 is missing the upstream direct repeat, as indicated by a bold line. SNPs in ser. Newport-III (sp9) and Typhimurium (sp12) are shown by variations in colour of the box.

Table 1. Number of alleles and number of spacers

Serovar	CRISPR1		CRISPR2	
	No. of alleles	No. of spacers*	No. of alleles	No. of spacers*
Enteritidis	7	7.9 (±1.7)	7	9.3 (±1.0)
Newport-II	9	21.8 (±3.8)	10	17.4 (±4.4)
Newport-III	14	11.5 (±4.3)	18	15.6 (±4.5)
Heidelberg	13	22.7 (±5.5)	8	15.1 (±2.2)
Typhimurium	18	15.8 (±7.6)	25	29.6 (±4.6)
Total	61		68	

*Values shown are the mean (±SD) number of spacers per array.

Serovars Newport-III, Typhiurium and Heidelberg also share four internal CRISPR2 spacers (sp8, 9, 13 and 14, with respect to ser. Newport-III). We also observed that although individual spacers are shared, their relative abundance within a serovar is often skewed. For example, ser. Typhimurium CRISPR2 sp31 is only found in two arrays (two isolates) but is present in all 13 ser. Heidelberg arrays (89 isolates). This spacer is also observed in 16/18 ser. Newport-III arrays (79/84 isolates).

Finally, there are no instances of spacer duplication among the ser. Heidelberg CRISPR arrays that we sequenced. Conversely, there are 11 duplicated spacers within our ser. Typhimurium isolates, including sp3 and 26 which are also found in ser. Heidelberg (Fig. 3), suggesting that different selective pressures exist on different serovars, driving the evolution of spacer content.

CRISPR array: last common ancestors (LCAs)

For each CRISPR locus in each serovar, we have indicated the LCA in Figs 2 and 3. Given that most differences arise from spacer loss or duplication, rather than acquisition, and that spacer order within an array is well maintained, we define the LCA as an array containing a full complement of spacers that are possible within a single serovar. With the exception of ser. Newport-II (CRISPR1 and 2) and Typhimurium (CRISPR2), an array identical to the LCA was observed within one or more of the *Salmonella* isolates screened.

Two distinct sets of CRISPR arrays in ser. Newport

Serovar Newport is polyphyletic, with three distinct lineages (Sangal *et al.*, 2010). As previously demonstrated (Fabre *et al.*, 2012), we were able to identify two of these (Lineage II and III) by CRISPR sequence analysis (Figs 2 and 3). There are no shared spacers among either of the lineages. Unexpectedly, we identified two isolates that each bear a Newport-III CRISPR1 locus but have a Newport-II CRISPR2 locus (Fig. S1). We note that both CRISPR1 arrays are different in the two isolates, as are the CRISPR2 arrays, confirming that these are distinct ser. Newport strains.

Direct repeat polymorphisms

While analysing the CRISPR array sequences, we noticed that many direct repeat variants (DRVs) exist that typically contain one or two SNPs, or small deletions, with respect to the consensus sequence. We identified 21 variants: 16 with one SNP, four with two SNPs and one with a single base deletion (Fig. 4a). The CRISPR1 locus of ser. Enteritidis is most highly conserved as all but one direct repeat (in array 66, present in only one isolate) have the consensus sequence (Fig. 4b). In contrast, the CRISPR2 loci have three distinct DRVs. The highest number of DRVs seen in a single locus was in ser. Typhimurium CRISPR2 (6/37 direct repeats). There does not seem to be a bias toward frequency of DRVs in one CRISPR locus versus the other (Fig. S2).

We next wanted to determine whether these DRVs were conserved, specifically whether they were associated with the same spacer(s) and whether they existed in distinct serovars. Regarding the former, we observed that DRVs were always associated upstream of the same spacer(s) with a single exception, DRV3 (Fig. S2). This variant is found next to the same spacer that is present in CRISPR1 of ser. Typhimurium and Heidelberg (sp18 and sp17, respectively) but is also seen with the leader proximal spacer of the same locus in ser. Typhimurium (sp32). There is no sequence similarity between these spacers and we assume that the SNP responsible for the DRV occurred independently. If a spacer is present in more than one serovar, the cognate DRV is also present, as demonstrated in ser. Typhimurium and Heidelberg. Otherwise, DRVs are not shared among the four serovars.

We found two examples where a DRV/spacer association was not conserved: (i) DRV7 occurs upstream of sp11 in ser. Newport CRISPR2 array 145, but not in other arrays also containing this spacer; and (ii) DRV21, which is upstream of sp7 in CRISPR1 array 66 (ser. Enteritidis), is not found in the related array 1.

Conservation of *cas* genes within a serovar

To study the diversity of the eight *cas* genes, we extracted and aligned these sequences from 206 genome assemblies.

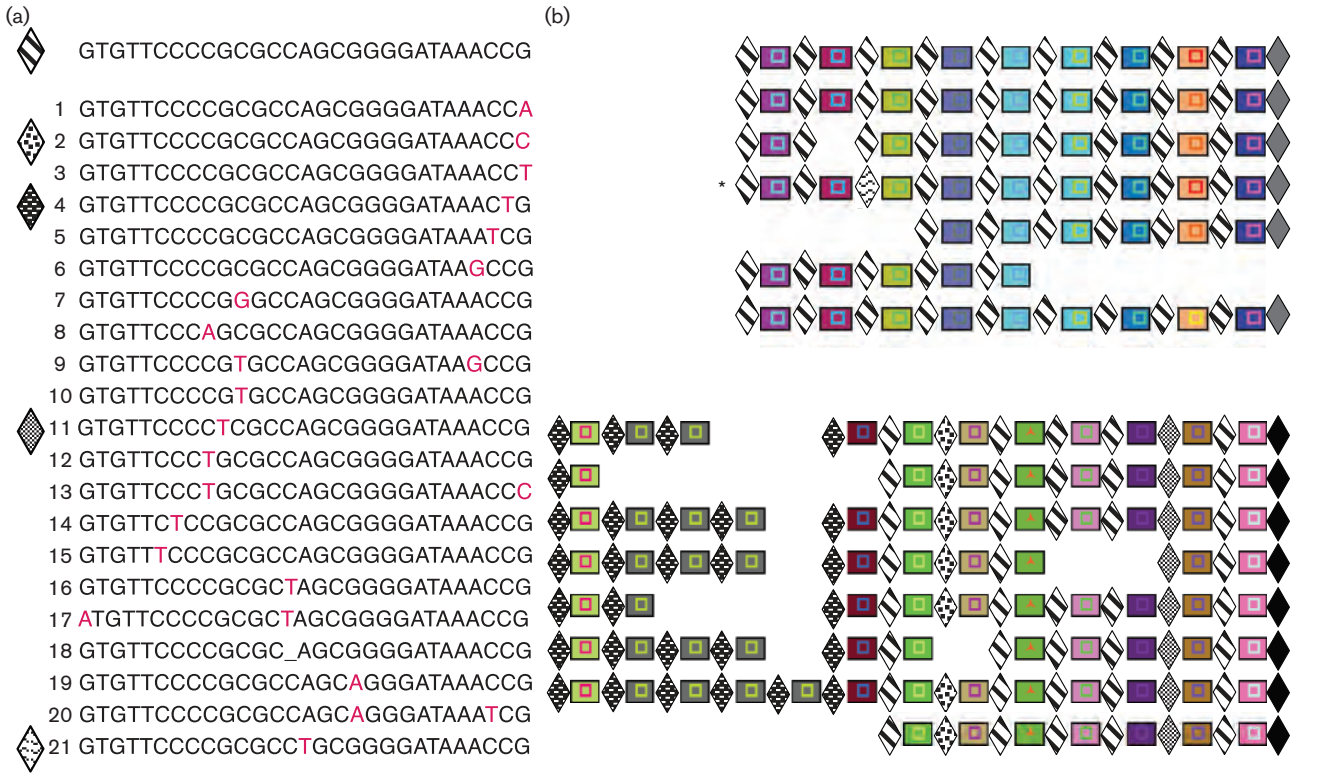

Fig. 4. Analysis of DRVs. (a) List of all DRVs identified in this study. The top sequence is the consensus sequence. (b) DRVs in ser. Enteritidis CRISPR1 (top) and CRISPR2 (bottom) arrays. A specific variant is always associated with the same spacer(s); where a spacer is missing, its cognate variant is also missing. A single array (*), occurring in only one isolate, contains an SNP resulting in a DRV.

In all but four cases, the predominant allele was observed in >91 % of the isolates within a serovar (Fig. 5). In 75 % of cases (30/40), all isolates within a serovar had identical *cas* alleles (see dark blue boxes in Fig. 5). Nucleotide alignments of the entire *cas* operon showed high conservation within a serovar (>99.9 % identity). Most strikingly, there is 100 % nucleotide identity among all 46 ser. Heidelberg isolates we analysed. Serovar Newport-II has the least conserved *cas* operon, as only five *cas* genes (*cse2*, *cas7*, *cas6*, *cas1* and *cas2*) are 100 % identical in all isolates. The *cas* genes that differ within a serovar generally arise from the presence of a single SNP.

Among all serovars, the two most conserved individual genes were *cas2* and *cse2*, for which only five alleles for each were identified (i.e. all isolates of each serovar contained the same allele). Interestingly, comparative analysis across the different serovars shows that *cas2* has a high level of nucleotide identity (98.30 %) but *cse2* has the lowest (83.42 %; Fig. 5). Although we identified ten distinct *cas3* alleles, these did not differ much at the nucleotide level (98.76 % nucleotide identity across four serovars, excluding ser. Newport-II isolates). Additionally, compared with the other serovars the *cas3* gene is in the reverse orientation in ser. Newport-II isolates and is separated from *cse1* by 357 nt (Fig. S3).

Differences in the *cas* operon among different serovars

In addition to determining the nucleotide identity, we wanted to visualize the differences between *cas* genes across the four different serovars. We aligned the sequences of the predominant *cas* operon from each serovar to each other (the isolates from which these sequences were extracted are indicated in Table S2). For this analysis, we did not include ser. Newport-II as the *cas* genes from this serovar have already been shown to be very distinct from those of the remaining serovars under investigation here (Pettengill *et al.*, 2014; Timme *et al.*, 2013). We used ser. Typhimurium as a reference, annotating SNPs with respect to this *cas* operon. We made four observations: first, there are *cas* sequences that are shared between serovars (Fig. 6). Specifically, there are three genes, *cse2*, *cas6e* and *cas2*, that are identical at the nucleotide level between ser. Typhimurium and Heidelberg. In addition, between these two serovars, there are only two SNPs (one synonymous, one non-synonymous) in *cas7* and one SNP (non-synonymous) in *cas1*. These observations also reflect the similarities seen in spacer composition of the CRISPR arrays. Second, there are several SNPs shared between the different serovars. For example, six of eight SNPs in *cas3* of ser. Enteritidis are also

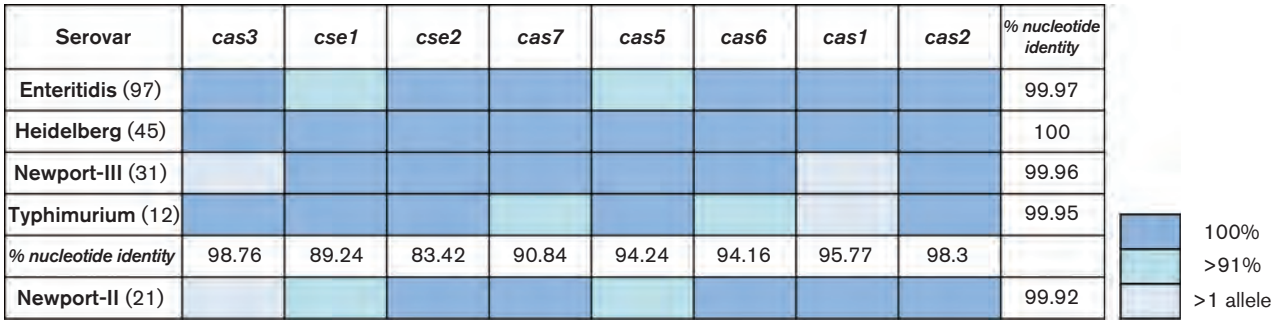

Fig. 5. Conservation of *cas* genes within a serovar. Matrix showing the conservation of *cas* genes within a serovar and of individual *cas* genes across different serovars. The number of isolates of each serovar is shown in parentheses and the nucleotide identity, as determined by the number of SNPs present, is shown in the right-hand column. Dark blue boxes represent 100 % sequence identity within a serovar and light blue boxes represent >91 % sequence identity within a serovar. Grey boxes represent the presence of two predominant alleles for an individual *cas* gene within a serovar.

found in ser. Heidelberg. Also in *cas3*, eight SNPs in ser. Newport-III are shared with ser. Heidelberg and two SNPs are common to ser. Newport-III, Enteritidis and Heidelberg. Third, *cas2* is the most conserved gene across the different serovars; there are no SNPs in ser. Heidelberg with respect to ser. Typhimurium, and although ser. Enteritidis and Newport-III contain one and five SNPs, respectively, all are synonymous.

Finally, unlike most of the *cas* genes, *cas5* differs dramatically between ser. Typhimurium and Heidelberg. Instead the majority of SNPs in ser. Heidelberg are shared with ser. Enteritidis, suggesting a horizontal gene transfer event. Also indicative of a horizontal gene transfer is the presence of numerous SNPs in the *cas* operon of ser. Newport-III, specifically between the 3′ end of *cas3* and *cas7*.

Conservation of leader sequences

We extracted both CRISPR1 and CRISPR2 leader sequences from our whole genome assemblies and aligned them according to serovar. Within a serovar, all leaders were identical for both CRISPRs with the exception of a single SNP in one CRISPR2 leader from a ser. Newport-II isolate (isolate SEEN443). Furthermore, for CRISPR1, ser. Typhimurium and Heidelberg shared the same leader sequence (Fig. S4). Serovars Enteritidis and Newport-III have one and two SNPs, respectively, compared with ser. Typhimurium and Heidelberg. For CRISPR2, ser. Typhimurium, Heidelberg and Newport-III all have the same leader sequence and this differs from ser. Enteritidis leaders by four SNPs and also from ser. Newport-II by four SNPs. The Newport-II CRISPR1 leader is divergent from the consensus CRISPR1 leader sequences but shares similarities with both CRISPR1 and CRISPR2 leader sequences.

Identification of phage/plasmid protospacers

Given the established immune function of CRISPR-Cas systems, we next sought to determine whether any *Salmonella* spacer matches phage or plasmid sequences. We used CRISPRTarget to identify possible protospacers (Biswas *et al.*, 2013) and defined a match as five or fewer SNPs (84 % match or ⩾27/32 nt) between spacer and protospacer (Table S3). Among 800 arrays analysed from 400 isolates, we identified 179 unique spacers for which we found putative protospacer matches for only one-quarter (42/179) (Fig. 7). Of these, 19 (10 % of the total) were found in phage or prophage sequences and only three

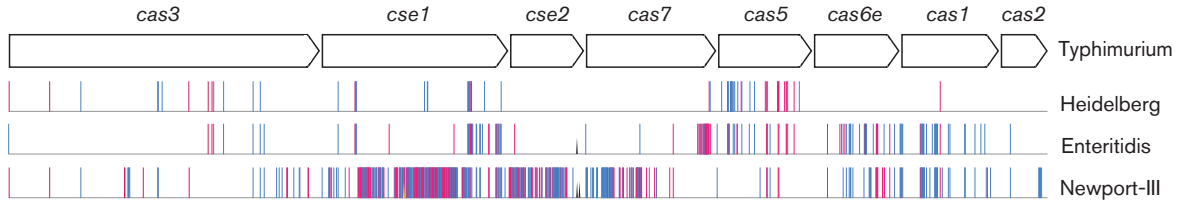

Fig. 6. Conservation of *cas* genes across different serovars. The sequences of the predominant *cas* operon for each serovar were aligned with respect to ser. Typhimurium. Synonymous SNPs and non-synonymous SNPs are indicated in blue and pink, respectively. The black arrowheads in *cse2* correspond to small deletions and the yellow arrowhead in *cse1* corresponds to an insertion. All maintain the reading frame.

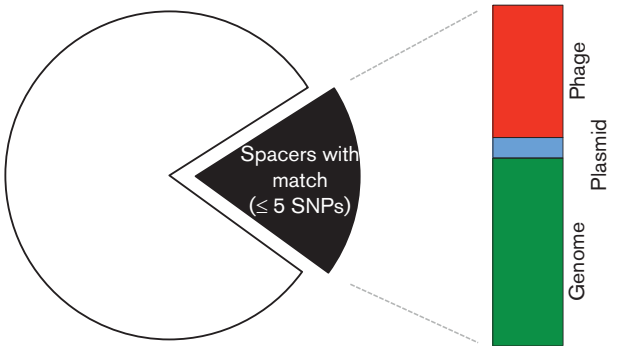

Fig. 7. Distribution of protospacers. Pie chart showing the number of unique spacers with potential protospacer matches to fewer than six SNPs (or ⩾27/32 nt matching). The distribution of these protospacers is shown to the right.

(2 %) protospacers were found in plasmid sequences. Somewhat surprisingly, 27 (15 %) protospacers were found in bacterial genomes (Fig. S5). These were in regions distinct from any prophage sequences and approximately half match to *Enterobacteriaceae* genomes, the most frequent being *Salmonella*, *E. coli*, *Klebsiella* sp. and *Erwinia* sp. There were five cases where a single spacer had protospacers within more than a single element. For two of these (ser. Heidelberg CRISPR1 sp29 and ser. Typhimurium CRISPR2 sp26), protospacers were found within a prophage, plasmid and a genome, although none of the genome protospacers was within *Salmonella*. The three remaining spacers all had protospacers in both prophages or phages and bacterial genomes (distinct from a prophage).

DISCUSSION

The overarching goal of this study was to provide an in-depth sequence analysis and characterization of the type I-E CRISPR-Cas system in *Salmonella*. Although closely related to the type I-E of *E. coli* (Touchon & Rocha, 2010), there are some important differences. For example, all *Salmonella* analysed to date exclusively harbour a type I-E system, whereas some *E. coli* have been shown to contain type I-F systems (Díez-Villaseñor *et al.*, 2010). Although there are some similarities in regulation of the *cas* operon between *E. coli* and *Salmonella* ser. Typhi (Medina-Aparicio *et al.*, 2011; Westra *et al.*, 2010), in *E. coli cas3* is transcribed independently from the remaining *cas* genes as there is an intergenic region between *cas3* and *cse1* that functions as a promoter (Pul *et al.*, 2010). In *Salmonella*, this intergenic region does not exist (there is no sequence similarity between this region and the 357 nt sequence in ser. Newport-II isolates). We analysed the three functional elements that comprise CRISPR-Cas: the *cas* genes, the leader sequence and the CRISPR array. While other research groups have studied *Salmonella* CRISPR loci (Fabre *et al.*, 2012; Liu *et al.*, 2011a; Pettengill *et al.*, 2014; Timme *et al.*, 2013), the entire CRISPR-Cas system has not been previously evaluated in such a large collection.

Our data show that at the serovar level all CRISPR-Cas elements are extremely well conserved. In addition to the similarities between ser. Typhimurium and Heidelberg, comparison of leader sequences and *cas* genes across all four serovars highlights a high level of conservation. Our study presents some novel insight into *Salmonella* CRISPR evolution with respect to leader sequences and organization of direct repeats within the arrays. As shown by others, we confirm that new spacers do not seem to be acquired by *Salmonella*, especially given that these isolates were collected over a 5-year period from distinct locations. In addition, we provide a comprehensive analysis of protospacer identification.

We identified 129 distinct CRISPR arrays, 61 for CRISPR1 and 68 for CRISPR2, in total and these contain 179 unique spacers. From a serotyping perspective, identification of spacers that are unique to a given serovar can be useful for designing high-throughput serovar-specific assays (Fabre *et al.*, 2012). In an immune active system, array differences arise from spacer acquisition (Tyson & Banfield, 2008). As shown in this study and others (Fabre *et al.*, 2012; Liu *et al.*, 2011a; Shariat *et al.*, 2013a), the majority of polymorphisms in *Salmonella* CRISPR arrays exist as a result of deletion of spacer-repeat units and this seems to occur most commonly with internal spacers. The low number of arrays missing the first or last spacer suggests some selection toward maintenance of these spacers and perhaps integrity of the array. Beyond this there is no selection for any particular region of the array from which spacers are lost. We specifically note that although spacers are lost, this occurs within the context of a spacer and its cognate direct repeat, thus maintaining the integrity of the array. This organization probably results from homologous recombination at the direct repeat sequence, thus maintaining the integrity. Such maintenance may have important implications if the CRISPR arrays provide an, as yet undetermined, alternative function that may require mature crRNAs.

CRISPR-Cas activity can be defined by acquisition of new spacers, transcription and processing of the mature crRNAs, or by interference. We found several lines of evidence that support greater activity and maintenance of CRISPR1 versus CRISPR2. First, ser. Heidelberg does not contain any spacers in CRISPR2 that are not found in ser. Typhimurium whereas its CRISPR1 locus contains seven spacers that are not found in ser. Typhimurium, suggesting that CRISPR1 is the more recently active of the two loci. Second, the presence of unique spacers at the leader proximal end of the array in two different ser. Typhimurium CRISPR1 alleles suggests that these have been recently acquired; we did not see any such unique spacers in any CRISPR2 loci. Third, although spacer loss happens in both arrays, spacer duplication exclusively occurs in CRISPR2. Additionally, there are two instances where a single spacer has been duplicated multiple times in CRISPR2. This suggests that there might be stronger selective pressure on maintaining the integrity of the CRISPR1

locus. Finally, of the 19 protospacers that are in phage or prophage regions, 15 match to spacers in CRISPR1.

There are two ser. Typhimurium CRISPR arrays with 'new' spacers at the leader proximal end. It is interesting to note that one of these, array 143, is also the most well-maintained ser. Typhimurium CRISPR1 locus as it contains a full complement of 32 spacers, with respect to the LCA. This implies that this particular isolate perhaps has a more active CRISPR system than the others we analysed. As our analysis showed that sp32 in array 143 appears to be self-targeting, we sequenced the protospacer region in the same isolate and found a 100 % match within a lipid kinase gene (data not shown). The absence of a correct protospacer adjacent motif upstream of the protospacer, plus the presence of a DRV directly upstream of the spacer (see below), probably impedes self-targeted CRISPR-Cas interference (Stern *et al.*, 2010). We were unable to identify a protospacer that matched the new spacer in array 134. Self-targeting spacers have been seen before and usually cluster at the leader proximal end of the array, presumably because if they target self they would not be maintained within a mature array. When self-targeting spacers are observed, they are often associated with an inactivating feature such as an improper protospacer adjacent motif or by mutations/loss of one or more *cas* genes (Stern *et al.*, 2010). In our dataset, we identify five self-targeting spacers with a perfect (32/32 bp) nucleotide match to *Salmonella* genomes (Fig. S5). Two of the five are associated with DRVs, which may affect processing of the pre-crRNA. A third self-targeting spacer, sp28 in ser. Typhimurium (Fig. S2), is missing the direct repeat downstream, and thus will not be cleaved by Cas6. Of the other two self-targeting spacers, one is in a CRISPR2 locus (ser. Typhimurium) and the other is within the older portion of a CRISPR1 locus (ser. Newport-III). This observation suggests that the arrays are, or were, active and that abrogation of self-targeting to the genome was promoted by removal or mutation of the direct repeat.

Although first suggested by Grissa *et al.* (2007), it was subsequently demonstrated that the leader proximal direct repeat is used as a template when a new direct repeat–spacer cassette is added during the acquisition process (Yosef *et al.*, 2012). We can see an excellent example of this in the CRISPR2 locus of ser. Enteritidis (Fig. 4b). According to their model, the SNPs that define DRV11 and DRV2 occurred after acquisition of sp2 and sp6, respectively. The SNP that is responsible for DRV5 probably occurred after addition of sp8 and was used as a template for addition of subsequent sp9 and sp10. Given this, DRV3, which lies upstream of the CRISPR2 leader proximal spacer in ser. Typhimurium, would be expected to be maintained upon addition of a new spacer.

Regarding the *cas* genes, there is a remarkable level of conservation both within a locus and across the four serovars that we examined here. Within a serovar [considering ser. Newport-II and Newport-III as different serovars due to their polyphyletic nature (Sangal *et al.*, 2010)], there is generally a single predominant allele for each *cas* gene. In ser. Heidelberg, with the exception of *cas3*, there was a single allele for each *cas* gene that was present in all 46 genome sequences that we interrogated (Fig. 5a). Comparison of *cas* gene sequences across serovars shows that there is a significant amount of conservation. For example, three *cas* genes are 100 % identical between ser. Typhimurium and Heidelberg and two others, *cas1* and *cas7*, have one and two SNPs, respectively. For our analysis, we chose to use ser. Typhimurium sequences as a reference; while SNPs exist with respect to this reference, several SNPs are shared between at least two of the three other serovars, for example in *cas3*, *cas5*, *cas6e* and *cas1*. Given that the *cas* operon is ~8.5 kb in length and the established divergence of these different serovars, this level of sequence identity is remarkable.

In speculating whether the *Salmonella* CRISPR-Cas system provides immunity, our data are similar to observations made within *E. coli*, where the CRISPR system does not exhibit typical characteristics of an active immune defence system (Touchon *et al.* 2011). However, our data provide somewhat of a conundrum: some evidence demonstrates a putative immune function, reflecting historical activity, while other data show lack of proposed CRISPR activity, instead reflecting current inactivity and perhaps a transition to a new functional role. All three elements are conserved: within a serovar, the nucleotide identity over the ~8.5 kb *cas* operon is >99.9 %, the leader sequences are identical and the CRISPR arrays are also conserved (notwithstanding spacer duplication and loss, there are few, if any, SNP occurrences within the arrays themselves). Specifically, the repeat–spacer–repeat integrity is maintained and self-targeting spacers are associated with DRVs. Across serovars (except ser. Heidelberg and Typhimurium CRISPR2), the spacer composition is different, as would be expected from an active immune system. Conversely, our data bolster the hypothesis that *Salmonella* CRISPR-Cas were historically active and are now evolving toward a CRISPR-Cas system with minimal immune activity: we do not see many instances of spacer acquisition except for the two putative acquisition events in ser. Typhimurium, and only a minority (12 %) of the total spacers show protospacer matches in mobile genetic elements. With one exception (1/19 spacers; ser. Newport-III CRISPR1 sp. 9), all spacers that have phage matches also have protospacer matches within prophages, providing evidence of an inactive or inept immune system, given that these viruses were able to integrate into the *Salmonella* genome. By comparison, in an active CRISPR-Cas system such as that of *Streptococcus thermophilus*, 77 % of spacers have viral protospacer matches (Horvath *et al.*, 2008). Spacers of bacterial origin have been observed in other bacteria, for example in *Yersinia pestis* (Riehm *et al.*, 2012). It is also interesting to note the imbalance of spacer maintenance where identical spacers are present in more than one serovar. It is tempting to speculate this was caused by an

immune-driven functional selection in one serovar versus another. However, given what we have observed of the *Salmonella* CRISPR-Cas system, it is more likely that loss of these particular spacers occurred soon after serovar divergence in one or more isolates which subsequently propagated, and thus are absent in a larger number of contemporary isolates. Finally, in all ser. Newport-II isolates, the *cas3* sequence is encoded in the opposite orientation; further work is required to determine the functional significance of this.

Protospacers within prophage regions have been observed in other bacteria; extensive spacer matches to temperate phages have been observed in *Pseudomonas aeruginosa* and *Streptococcus pyogenes* (Cady *et al.*, 2011; Deltcheva *et al.*, 2011). In the former, CRISPR-Cas has been linked to the regulation of biofilm formation (Zegans *et al.*, 2009). In other examples, protospacers in prophages have also been identified in *Clostridium difficile* (Hargreaves *et al.*, 2014), and recent work in *Staphylococcus epidermidis* shows that spacers matching to prophage regions can tolerate lysogeny but target the virus upon viral induction (Goldberg *et al.*, 2014).

We have provided a thorough characterization of *Salmonella* CRISPR-Cas systems in four prevalent clinical serovars. Our findings suggest that from an immune perspective, *Salmonella* CRISPR-Cas was at one point active, but is no longer so. However, the conservation of their components, both within a serovar and across divergent serovars, indicates these loci may have an alternative yet highly conserved function in *Salmonella*.

It is becoming apparent that CRISPR-Cas systems do have alternative functions (Bondy-Denomy & Davidson, 2014; Westra *et al.*, 2014). For example, these systems have been shown to be involved in biofilm formation (Zegans *et al.*, 2009), host infection in humans and amoeba (Gunderson & Cianciotto, 2013; Sampson *et al.*, 2013), symbiotic colonization in nematodes (Veesenmeyer *et al.*, 2014) and DNA damage (Babu *et al.*, 2011). If an alternative function exists in *Salmonella* and is potentially driven, at least in part, by complementarity between a crRNA and its genetic target, our finding that 15 % of spacers target bacterial genomes and that nearly one-fifth of these protospacers are within *Salmonella* genomes supports this hypothesis.

ACKNOWLEDGEMENTS

We thank Dr Jennifer Mandel for helpful discussions and Dr Phillipe Horvath for sharing the macro (CRISPR DB II Excel Macro). This study was supported by a United States Army Research Office grant to E. G. D. (W911NF-11-1-0442).

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G. & other authors (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79**, 484–502.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712.

Bhaya, D., Davison, M. & Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45**, 273–297.

Biswas, A., Gagnon, J. N., Brouns, S. J., Fineran, P. C. & Brown, C. M. (2013). CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol* **10**, 817–827.

Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561.

Bondy-Denomy, J. & Davidson, A. R. (2014). To acquire or resist: the complex biological effects of CRISPR-Cas systems. *Trends Microbiol* **22**, 218–225.

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964.

Cady, K. C., White, A. S., Hammond, J. H., Abendroth, M. D., Karthikeyan, R. S., Lalitha, P., Zegans, M. E. & O'Toole, G. A. (2011). Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. *Microbiology* **157**, 430–437.

Centers for Disease Control & Prevention. (2011). *National Salmonella Surveillance Annual Report 2011*. Atlanta, GA: CDC.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J. & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P. & Moineau, S. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1390–1400.

Díez-Villaseñor, C., Almendros, C., García-Martínez, J. & Mojica, F. J. M. (2010). Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* **156**, 1351–1361.

DiMarzio, M. J., Shariat, N., Kariyawasam, S., Barrangou, R. & Dudley, E. G. (2013). Antibiotic resistance in *Salmonella* Typhimurium associates with CRISPR sequence type. *Antimicrob Agents Chemother* **57**, 4282–4289.

Fabre, L., Zhang, J., Guigon, G., Le Hello, S., Guibert, V., Accou-Demartin, M., de Romans, S., Lim, C., Roux, C. & other authors (2012). CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS ONE* **7**, e36995.

Fricke, W. F., Mammel, M. K., McDermott, P. F., Tartera, C., White, D. G., Leclerc, J. E., Ravel, J. & Cebula, T. A. (2011). Comparative genomics of 28 *Salmonella enterica* isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *J Bacteriol* **193**, 3556–3568.

Garneau, J. E., Dupuis, M. È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H. & Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71.

Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **514**, 633–637.

Grimont, P. & Weill, F. (2007). *Antigenic Formulae of the Salmonella Serovars*, 9th edn. Geneva: WHO.

Grissa, I., Vergnaud, G. & Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

Gunderson, F. F. & Cianciotto, N. P. (2013). The CRISPR-associated gene *cas2* of *Legionella pneumophila* is required for intracellular infection of amoebae. *mBio* **4**, e00074-13.

Hale, C., Kleppe, K., Terns, R. M. & Terns, M. P. (2008). Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* **14**, 2572–2579.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. & Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956.

Hargreaves, K. R., Flores, C. O., Lawley, T. D. & Clokie, M. R. (2014). Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. *MBio* **5**, e01045-13.

Horvath, P., Romero, D. A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. & Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1401–1412.

Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565–1575.

Lillestøl, R. K., Redder, P., Garrett, R. A. & Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59–72.

Liu, F., Barrangou, R., Gerner-Smidt, P., Ribot, E. M., Knabel, S. J. & Dudley, E. G. (2011a). Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. *Appl Environ Microbiol* **77**, 1946–1956.

Liu, F., Kariyawasam, S., Jayarao, B. M., Barrangou, R., Gerner-Smidt, P., Ribot, E. M., Knabel, S. J. & Dudley, E. G. (2011b). Subtyping *Salmonella enterica* serovar *enteritidis* isolates from different sources by using sequence typing based on virulence genes and clustered regularly interspaced short palindromic repeats (CRISPRs). *Appl Environ Microbiol* **77**, 4520–4526.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I. & other authors (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467–477.

Medina-Aparicio, L., Rebollar-Flores, J. E., Gallego-Hernández, A. L., Vázquez, A., Olvera, L., Gutiérrez-Ríos, R. M., Calva, E. & Hernández-Lucas, I. (2011). The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* serovar Typhi. *J Bacteriol* **193**, 2396–2407.

Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174–182.

Pettengill, J. B., Timme, R. E., Barrangou, R., Toro, M., Allard, M. W., Strain, E., Musser, S. M. & Brown, E. W. (2014). The evolutionary history and diagnostic utility of the CRISPR-Cas system within *Salmonella enterica* ssp. *enterica*. *PeerJ* **2**, e340.

Pourcel, C., Salvignol, G. & Vergnaud, G. (2005). CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663.

Pul, Ü., Wurm, R., Arslan, Z., Geißen, R., Hofmann, N. & Wagner, R. (2010). Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol* **75**, 1495–1512.

Riehm, J. M., Vergnaud, G., Kiefer, D., Damdindorj, T., Dashdavaa, O., Khurelsukh, T., Zöller, L., Wölfel, R., Le Flèche, P. & Scholz, H. C. (2012). *Yersinia pestis* lineages in Mongolia. *PLoS ONE* **7**, e30624.

Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y.-L. & Weiss, D. S. (2013). A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**, 254–257.

Sangal, V., Harbottle, H., Mazzoni, C. J., Helmuth, R., Guerra, B., Didelot, X., Paglietti, B., Rabsch, W., Brisse, S. & other authors (2010). Evolution and population structure of *Salmonella enterica* serovar Newport. *J Bacteriol* **192**, 6465–6476.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M.-A., Roy, S. L., Jones, J. L. & Griffin, P. M. (2011). Foodborne illness acquired in the United States–major pathogens. *Emerg Infect Dis* **17**, 7–15.

Shariat, N., DiMarzio, M. J., Yin, S., Dettinger, L., Sandt, C. H., Lute, J. R., Barrangou, R. & Dudley, E. G. (2013a). The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of *Salmonella enterica* subsp. *enterica* serovar Enteritidis. *Food Microbiol* **34**, 164–173.

Shariat, N., Kirchner, M. K., Sandt, C. H., Trees, E., Barrangou, R. & Dudley, E. G. (2013b). Subtyping of *Salmonella enterica* serovar Newport outbreak isolates by CRISPR-MVLST and determination of the relationship between CRISPR-MVLST and PFGE results. *J Clin Microbiol* **51**, 2328–2336.

Shariat, N., Sandt, C. H., DiMarzio, M. J., Barrangou, R. & Dudley, E. G. (2013c). CRISPR-MVLST subtyping of *Salmonella enterica* subsp. *enterica* serovars Typhimurium and Heidelberg and application in identifying outbreak isolates. *BMC Microbiol* **13**, 254.

Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* **26**, 335–340.

Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnes, C., Van Kessel, J. S., Karns, J. S., Musser, S. M. & Brown, E. W. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol* **5**, 2109–2123.

Touchon, M. & Rocha, E. P. C. (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE* **5**, e11126.

Touchon, M., Charpentier, S., Clermont, O., Rocha, E. P., Denamur, E. & Branger, C. (2011). CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J Bacteriol* **193**, 2460–2467.

Tyson, G. W. & Banfield, J. F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**, 200–207.

Veesenmeyer, J. L., Andersen, A. W., Lu, X., Hussa, E. A., Murfin, K. E., Chaston, J. M., Dillman, A. R., Wassarman, K. M., Sternberg, P. W. & Goodrich-Blair, H. (2014). NilD CRISPR RNA contributes to *Xenorhabdus nematophila* colonization of symbiotic host nematodes. *Mol Microbiol* **93**, 1026–1042.

Westra, E. R., Pul, Ü., Heidrich, N., Jore, M. M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M. & other authors (2010). H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. *Mol Microbiol* **77**, 1380–1393.

Westra, E. R., Buckling, A. & Fineran, P. C. (2014). CRISPR-Cas systems: beyond adaptive immunity. *Nat Rev Microbiol* **12**, 317–326.

Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338.

Yosef, I., Goren, M. G. & Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res* **40**, 5569–5576.

Zegans, M. E., Wagner, J. C., Cady, K. C., Murphy, D. M., Hammond, J. H. & O'Toole, G. A. (2009). Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *J Bacteriol* **191**, 210–219.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res* **39** (Suppl 2), W347-W352.

Edited by: R. Lan